
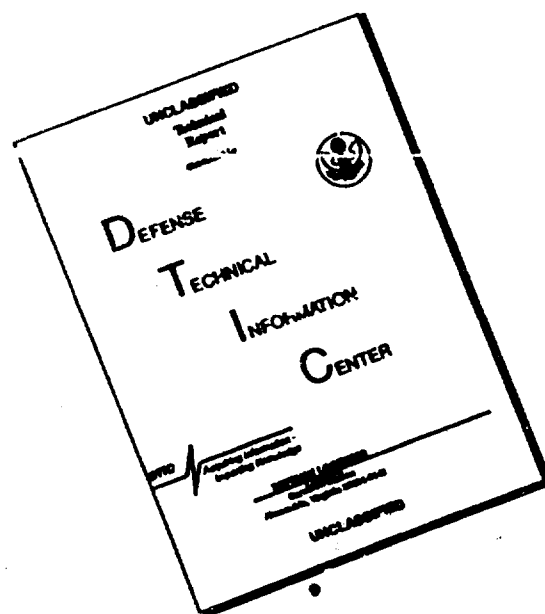


REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
1a. REPORT SECURITY CLASSIFICATION Unclassified		1b. RESTRICTIVE MARKINGS			
2a. SECURITY CLASSIFICATION AUTHORITY AD-A240 195 		3. DATE SEP 09 1991		4. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.	
5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 91 0714		6a. ADDRESS (City, State, and ZIP Code) Princeton University			
6b. ADDRESS (City, State, and ZIP Code) Department of Chemistry/Dept. of MAE Princeton, NJ 08544-1009		7a. NAME OF MONITORING ORGANIZATION AFOSR/NA			
7b. ADDRESS (City, State, and ZIP Code) Building 410, Bolling AFB DC 20332-6448		8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR/NA			
8b. OFFICE SYMBOL (if applicable) NA		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER AFOSR-89-0070			
10. SOURCE OF FUNDING NUMBERS		11. TITLE (Include Security Classification) (U) A Systematic Approach to Combustion Model Reduction and Lumping			
12. PERSONAL AUTHOR(S) Herschel Rabitz and Fredrick Dryer		13a. TYPE OF REPORT Final Tech. Report			
13b. TIME COVERED FROM 12/88 TO 12/90		14. DATE OF REPORT (Year, Month, Day) 91, 8, 1		15. PAGE COUNT 459	
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)			
FIELD	GROUP	SUB-GROUP	combustion modelling, chemical kinetics, lumping, reduction, sensitivity analysis, Lie algebra techniques		
19. ABSTRACT (Continue on reverse if necessary and identify by block number)					
<p>This report summarizes research activities completed over the past two years in the general area of combustion model reduction and lumping. The purpose of the research was for the further development of practical techniques capable of rendering complex combustion-transport models to their physical essence for realistic computational execution. The research followed three avenues of approach: a) sensitivity analysis, b) linear projective transformations; c) Lie algebraic techniques. The diversity of approach was necessitated by the complexity of the problem and significant progress was made in each area. Specific conclusions were made concerning the likely next level of research developments needed to advance these tools to practical fruition.</p>					
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input checked="" type="checkbox"/> DTIC USERS		21. ABSTRACT SECURITY CLASSIFICATION Unclassified			
22a. NAME OF RESPONSIBLE INDIVIDUAL Julian M Tishkoff		22b. TELEPHONE (Include Area Code) (202) 767-0465		22c. OFFICE SYMBOL AFOSR/NA	



DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REF ID: A66842
 TABLE OF CONTENTS

I. Background	1
II. Summary of the Completed Research	2
A. Lumping and Reduction Based on Sensitivity Analysis Techniques	2
B. Linear Projective Transformations for Lumping	3
C. Lie Algebraic Techniques for Lumping	4
III. Specific Research Advances	5
A. Lumping and Reduction Based on Sensitivity Analysis Techniques	6
B. Linear Projective Transformations for Lumping	10
C. Lie Algebraic Techniques for Lumping	14
IV. Participating Professional Personnel	17
V. Presentations	17
VI. Inventions	17
References	18
Appendix A	19
Appendix B	48
Appendix C	96
Appendix D	125
Appendix E	182
Appendix F	226
Appendix G	253
Appendix H	271
Appendix I	290
Appendix J	305
Appendix K	340
Appendix L	369
Appendix M	404
Appendix N	422

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Avail and/or	Special
<div style="display: flex; align-items: center;"> <div style="border: 1px solid black; padding: 2px; margin-right: 5px;"> DIST A1 </div> <div style="border: 1px solid black; width: 100px; height: 100px; margin-left: 5px;"></div> </div>	



AIR FO -
 100-100000

A SYSTEMATIC APPROACH TO COMBUSTION MODEL REDUCTION AND LUMPING

I. Background.

Extensive effort over many years has gone into the development of combustion models with the long-range aim of executing them in a practical fashion for engineering combustor design. The overall problem breaks into two strongly coupled components involving fluid mechanics and chemical kinetics. From a modeling perspective the number of dependent variables essentially determines the computational difficulty and the number of reactive species involved is generally the key factor. Thus, there is an enormous impetus to arrive at practical, as well as accurate, models of the reactive-transport processes that are reduced to their essential structure. This goal has been a long standing one in the field and is of rising significance due to recent advances in computational engineering applications.

Formally, the topics of reduction and lumping of kinetic systems address the problems stated above. Unfortunately until now there has been little systematic guidance on how to take a given problem and reduce its complexity in a systematic manner. Empirical rate laws have been employed with limited success, and the traditional use of the steady state approximation is often of limited value. The present research is founded on the desire to systematically develop reduction and lumping tools for producing simplified chemical and transport models in different combustion and kinetic environments. Secondly, we desire to create constructive techniques for both assessing the degree to which a reactive mechanism may be lumped and

providing a concrete means for achieving that goal in favorable cases. Although much progress has been made and significant steps in these directions were successfully performed during the tenure of this grant, much still remains to be pursued.

II. Summary of the Completed Research

The terms lumping and reduction are used here to denote two distinct types of reactive-transport model simplification. Lumping refers to a contraction or possibly elimination in the number of dependent variables (i.e., chemical species) while reduction refers to all other simplifications in the coupled kinetic system (i.e., an elimination of insignificant reactive steps, etc.). In some cases lumping may result from a direct elimination of identified insignificant species while in other cases lumping may be achieved by the creation of accurate effective reactive mechanisms. This distinction alone generally calls for the use of different techniques to achieve the dual goals of lumping and reduction. Furthermore, the overall complexity of the problem has led us to pursue three distinct approaches. Each has its own merits and has been developed to differing degrees of achievement. A summary of each technique and their respective capabilities is given in this section, while in Section III, a synopsis of the specific projects is presented in the format of an abstract of each of the works.

A. LUMPING AND REDUCTION BASED ON SENSITIVITY ANALYSIS TECHNIQUES. Serious attempts at developing sensitivity analysis for combustion kinetics goes back some fifteen years, with much of the basic developments occurring at Princeton. In essence, sensitivity analysis provides a means for

quantitatively assessing the overall relationship between the pool of dependent and independent variables in a reactive-transport system. In the present context, this assessment is achieved by computing a family of sensitivity coefficients which are gradients relating one variable to another. Thus, as an auxiliary component to performing the modelling alone, separate codes have been written to efficiently compute this analysis information. Although partial derivative sensitivity coefficients are used as a quantitative measure of the variable relationships, the results actually can be interpreted as the response of the reactive-transport system to a perturbation of one of its variables. In the development of these tools, it was recognized early on that this perturbation-response relationship should contain valuable information for identifying the significant and insignificant portions of reactive models. This identification can be focussed on lumping, where the goal is to identify species playing insignificant roles, or on reduction, where a singling out of insignificant rate constants or transport coefficients is the objective. These techniques have now been implemented to a rather high level, with a number of cases showing the capability of achieving significant simplifications. An intriguing result observed during this development was the presence of scaling and self-similarity behavior amongst the sensitivity coefficients in strongly coupled exothermic combustion systems. It has been argued that the presence of this surprising system behavior is a strong indicator that lumping and reduction may be successfully achieved.

B. LINEAR PROJECTIVE TRANSFORMATIONS FOR LUMPING. The need for lumping of complex reactive systems occurs in other areas besides combustion, and this

problem was recognized many years ago in the chemical engineering community. Dating from the mid-1960s, a considerable effort has gone into the development of linear transformation techniques to project the set of chemical species into a lower dimensional space while still preserving its essential character. Almost all of the prior work focussed on linear kinetic systems for which this approach is almost a trivial exercise. The work at Princeton has put this theory on a rigorous foundation and, most importantly, it has extended applications to fully nonlinear chemical kinetic systems including the presence of transport (refs. 6-10 summarized in Section III). It was possible to establish the criteria for the existence of transformations which will achieve exact lumping in a given system. Although exact lumping is highly unlikely to occur in realistic problems, establishing the criteria for its existence provided an important step in developing an algorithm for finding lumping transformations that can approximate exact lumping to the desired level of accuracy. This work, carried out over the past six years, represents a milestone upon which to build an even more broadly applicable theory of lumping based on nonlinear transformations. Notwithstanding the latter need for further research, the linear lumping transformation techniques were developed into a well-defined algorithmic framework for application where appropriate.

C. LIE ALGEBRAIC TECHNIQUES FOR LUMPING. The sensitivity analysis techniques for lumping in paragraph A above are based on the notion of examining the response to infinitesimal disturbances of the reactive transport system. In a similar vein, the use of Lie algebraic methods is also based on considering the fundamental properties of the generators of

infinitesimal transformations upon a differential equation system. However, unlike sensitivity analysis, Lie algebraic techniques extend these transformations in a global manner for finite disturbances. In reality, lumping is a finite alteration of the combustion system, and, in the case of sensitivity analysis, the coefficients are used as a quantitative indicator of what finite changes to perform. In contrast, Lie algebraic techniques hold potential for explicitly tracing the infinitesimal alterations up to a specific finite level for practical applications. This is a very ambitious goal; however, it is important to pursue if for no other reason than the fundamental insight such an exercise provides. Among the three lines of approach, it is apparent that the Lie algebraic method is both the most ambitious and at the earliest stage of development. The most important result emanating from the Lie algebraic research consisted of an identification of the classes of transformations of a reactive system and their ability to preserve the topological nature of the evolving reactive flow. In a more practical vein, specific generators for Lie algebraic transformations were found which satisfied an imposed degree of accuracy. In the long term, this approach holds promise for providing fundamental insight into the ability to lump broad classes of systems and to achieve practical means for their success.

III. Specific Research Advances

The following material consists of abstracts of the particular research papers developed during the tenure of this grant. The papers are drawn together under headings following those listed in Section II above.

A. LUMPING AND REDUCTION BASED ON SENSITIVITY ANALYSIS TECHNIQUES

1. The Effects of Thermal Coupling and Diffusion of the Mechanism of H₂ Oxidation in Steady, Premixed Laminar Flames¹

The work considered the question of why steady premixed laminar flames can be successfully described by highly reduced models, whereas the underlying mechanism is inherently complex. The calculations were performed on H₂-air systems. Sensitivity functions were evaluated and studied for diffusion-free situations, both isothermal and adiabatic, as well as for steady premixed flames. In the diffusion-free cases most reactions of a 38-step mechanism were shown to be influential in a distinct fashion. The form of the sensitivity functions is, however, radically changed and rendered self-similar by simultaneous thermal coupling and diffusion that introduce strong nonlinear coupling among the variables. Due to self similarity, the mechanism can be reduced to 15 reactions while keeping the temperature profile and the mass fraction profiles of molecular species almost unchanged in flame calculations. Furthermore, there exists an invariant subspace in the space of kinetic parameters such that large parameter perturbations along any vector in this subspace result in relatively small changes in the computed flame properties. By giving mechanistic interpretation to such parameter perturbations, the model can be simplified in many ways. In particular, a sequence of models was constructed in a stoichiometric H₂-air flame problem that converge to a 9-step reduced mechanism with quasi steady state assumptions in radicals except H, thereby resulting in a two-step

quasi-global model. All these approximations are unfeasible without the presence of molecular and thermal diffusion.

2. Parametric Sensitivity and Self-similarity in Thermal Explosion Theory²

Relations between thermal runaway (also called parametric sensitivity) and self-similarity are studied. Both concepts are sensitivity-related but deal with system properties that are independent of the choice of particular parameters being perturbed. This independence is emphasized by proposing a new generalized condition for parametric sensitivity. Criticality is defined as the point in the parameter space where the trajectory exhibits maximum sensitivity to arbitrary, unstructured perturbations applied at the temperature maximum. The condition reduces to the analysis of eigenvalues of the Jacobian matrix. In addition to its conceptual generality, the new condition shows that there exists no critical Semenov number for some values of the other parameters. The sensitivity functions are shown to satisfy self-similarity relations if and only if the system exhibits critical or supercritical behavior. The onset of self-similarity is explained in terms of two properties of explosion systems, both related to parametric sensitivity. First, the temperature is the dominant variable, and any perturbation in the system affects the conversion mainly through the changes induced in the temperature. This coupling of the variables is shown by decomposing the sensitivity functions into direct and indirect terms. Second, after some induction period, the sensitivity equations are pseudo-homogeneous, i.e., the system becomes relatively insensitive to parameter perturbations applied at later stages of the reaction. The two

properties enable one to explain self-similarity of sensitivity functions observed in many explosion and combustion systems. Relations to earlier parametric sensitivity and self-similarity conditions are discussed.

3. A Combined Stability-sensitivity Analysis of Weak and Strong Reactions of Hydrogen/Oxygen Mixtures³

Stability and sensitivity analysis are used to examine the ignition/reaction characteristics of dilute hydrogen-oxygen mixtures. The analysis confirms the existence of two distinct regions of ignition and fast reaction previously labelled "weak" and "strong" ignition, both of which are located in the explosive pressure-temperature domain and separated by a region related to the "extended" classical second limit. The stability analysis is based on an eigenanalysis of the Green's function matrix of the governing kinetic equations. The magnitudes of the largest (and system controlling) eigenvalue allow the strengths of the two process to be quantified, giving a clear definition to the terms "weak" and "strong". The sensitivities of the largest eigenvalue to the reaction rate constants of the mechanism pinpoint the elementary steps controlling the two ignition processes and the subsequent reaction. The associated eigenvectors yield the direction of change in species concentrations and temperature during the course of reaction. These vectors are found to be nearly constant during the induction period of both "weak" and "strong" ignition, thus producing constant overall stoichiometric reactions. The subsequent reaction of major reactants associated with "weak" ignition also has a constant overall reaction

vector, although, different than that during the induction period. However, the vector describing the reaction of major reactants associated with "strong" ignition is found never to be constant, but continuously changing beyond the induction period.

4. On the Use of Green's Functions for the Analysis of Dynamic Couplings:

Some Examples from Chemical Kinetics and Quantum Dynamics

The utility of individual elements of Green's functions matrices, in the investigation of dynamic couplings, is illustrated by offering examples from linear and nonlinear kinetics and quantum dynamics. The concept of reduced Green's functions affords a detailed characterization of the actual pathways mediating these couplings. Self-similarity behavior between different elements of the Green's function matrix indicates the presence of strong coupling between different variables of the model. We investigate the structure of the entire Green's function matrix to examine such self-similarity behavior and other simplifying characteristics of concern for physical insight as well as for economic modeling of the dynamic systems. Global structure in the entire Green's function matrix may be used to reduce the complexity (number of dependent variables) in a model.

5. Sensitivity Analysis of a Steady-state, Premixed Laminar CO-H₂-O₂ Flame⁵

The direct and very efficient Newton method for obtaining sensitivities of two-point boundary value problems is utilized for detailed exploration of a reacting-diffusing CO+H₂+O₂ steady-state premixed laminar flame. Sensitivity coefficients and Green's functions

calculated for this system offer exhaustive characterization and new insights into the role of diffusion and exothermicity in carbon monoxide oxidation kinetics. In particular, the reactions of the hydroperoxy radical with hydrogen, oxygen and hydroxyl radicals are found to be extremely important at all temperatures in the fuel lean (40 torr) flame studied here. The diffusive mixing of chemical species from the low and high temperature portions of the flame and the large heats of reaction associated with the hydroperoxy radicals are found to be responsible for the increased importance of these reactions.

B. LINEAR PROJECTIVE TRANSFORMATIONS FOR LUMPING

6. General Analysis of Approximate Lumping in Chemical Kinetics⁶

A general analysis of approximate lumping based on linear transformations has been developed. This analysis can be applied to any reaction system with n species described by $dy/dt = f(y)$, where y is an n -dimensional vector in a desired region Ω and $f(y)$ is an arbitrary n -dimensional function vector. Here we have considered lumping by means of a rectangular constant matrix M (i.e., $\hat{y} = My$, where M is a row-full rank matrix and \hat{y} has dimension \hat{n} not larger than n). The observer theory initiated by Luenberger was formally employed to obtain the kinetic equations and discuss the properties of the approximately lumped system. The approximately lumped kinetic equations have the same form $d\hat{y}/dt = Mf(\bar{M}\hat{y})$ as that for the exactly lumped ones, but depend on the choice of the generalized inverse \bar{M} of M . The (1,2,3,4) inverse is a

good choice of the generalized inverse of M . The equations to determine the approximate lumping matrices M has been developed. These equations can be solved by iteration. An approach for choosing suitable initial iteration values of the equations has been illustrated in several examples.

7. A General Analysis of Exact Lumping in Chemical Kinetics⁷

A general analysis of exact lumping is presented. This analysis can be applied to any reaction system with n species described by a set of first order differential equations $dy/dt = f(y)$, where y is an n -dimensional vector, $f(y)$ is an arbitrary n -dimensional function vector. Here we consider lumping by means of an $\hat{n} \times n$ real constant matrix M with rank $\hat{n}(\hat{n} < n)$. It is found that a reaction system is exactly lumpable if and only if there exist nontrivial fixed invariant subspaces M of the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$, no matter what value y takes, and the corresponding eigenvalues are the same for $J^T(y)$ and $J^T(\bar{M}y)$. Here the rows of M are the basis vectors of M and \bar{M} is any generalized inverse of M satisfying $M\bar{M} = I_{\hat{n}}$ with $I_{\hat{n}}$ being the \hat{n} -identity matrix. The fixed invariant subspaces of $J^T(y)$ can be obtained either from the simultaneously invariant subspaces of all A_k , where the A_k 's form the basis of the decomposition of $J^T(y)$ or by determining the fixed $\text{Ker} \{ \Pi_1 (J^T(y) - \lambda_1 I_n)^{r_1} \Pi_j [\alpha_j^2 + r_j^2] I_n - 2\alpha_j J^T(y) + (J^T(y))^2]^{r_j} \}$, where λ_1 , $\alpha \pm ir_j$ are the real and nonreal eigenvalues of $J^T(y)$ and λ_1 , α_j and r_j are usually functions of y ; r_1 , r_j are nonnegative integers. The kinetic equations of the lumped system can be described as $d\hat{y}/dt = Mf(\bar{m}\hat{y})$. This method is illustrated by some simple examples.

8. The Determination of Constrained Lumping Schemes for a Reaction System in the Whole Composition Space⁸

Two new approaches to the determination of constrained lumping schemes have been developed. They are based on the property that the lumping schemes validated in the whole composition Y_n -space of y are only determined by the invariance of the subspace spanned by the row vectors of lumping matrix M with respect to the transpose of the Jacobian matrix $J^T(y)$ for the kinetic equations. We have proved that when a part of a lumping matrix M_G is given, each row of the part of the lumping matrix to be determined M_D is a certain linear combination of a set of eigenvectors of a special symmetric matrix. This symmetric matrix is related to M_G^T and $A_k M_G^T$, where A_k are the basis matrices of $J^T(y)$. It has been shown that the approximate lumping matrices containing M_G with different row number $\hat{n}(\hat{n} < n)$ and global minimum errors can be determined by an optimization method. Using the concept of the minimal invariant subspace of a constant matrix over a given subspace one can directly obtain the lumping matrices containing M_G with different \hat{n} . The accuracy of these lumping matrices was shown to be satisfactory in several sample calculations.

9. Determination of Constrained Lumping Schemes for Nonisothermal First-order Reaction Systems⁹

The direct approach to determining the constrained lumping schemes summarized in items 6-8 above has been applied to nonisothermal first-order reaction systems. The constant basis matrices of the transpose of the Jacobian matrix for the kinetic equations were replaced by a set of

rate constant matrices at different temperatures which properly cover the desired temperature region. This approach allows for the consideration of a distribution of temperatures as well as directly incorporating an energy balance equation. As an illustration, the technique was successfully applied on a model for petroleum cracking.

10. A General Lumping Analysis of a Reaction System Coupled with Diffusion¹⁰

A general lumping analysis of a reaction system coupled with diffusion is presented. This analysis can be applied to any reaction system with n species for both steady-state and transient conditions. Here we consider lumping by means of an $\hat{n} \times n$ constant matrix M with rank $\hat{n}(\hat{n} \leq n)$. When the diffusivity is independent of position and concentration vectors r and y , it is found that under steady-state conditions a reaction system having species concentration vector $y(r)$ coupled with diffusion is exactly lumpable if and only if there exist nontrivial fixed $J^T(y(r))D^{-1}$ invariant subspaces M (here $J^T(y(r))$ is the transpose of the Jacobian matrix for the chemical reaction rate vector $f(y(r))$ and D^{-1} is the inverse of the constant effective diffusivity matrix), no matter what value $y(r)$ takes; under transient conditions there exist simultaneously D - and $J^T(y(r,t))$ -invariant subspaces M . When D is a function of position or concentrations, M is simultaneously invariant to $J^T(y)$ and $D(r)$, $D(y(r,t))$. The same approach to determine the constrained approximate lumping schemes for a non-diffusion system can be used in a reaction-diffusion one except that the constant basis matrices A_k 's of $J^T(y)$ are replaced by $B_k = A_k D^{-1}$ under steady-state conditions or the extra matrix D is added under transient conditions.

For nonconstant D , the basis constant matrices D_1 's of $D(r)$, $D(y(r))$ or $D(y(r,t))$ are added.

C. LIE ALGEBRAIC TECHNIQUES FOR LUMPING

11. Lie Algebraic Factorization of Multivariable Evolution Operators:
Convergence Theorems for the Canonical Case¹¹

This work is devoted to establishing the convergence theorems for the canonical case of the Lie algebraic factorization of multivariable evolution operators. The definition and various properties of $\bar{\xi}$ -approximants are given in a companion paper. The theorems presented in this paper give some sufficient conditions for the convergence of the $\bar{\xi}$ -approximant sequences. Proofs are given for a specific region of the variables space appearing in the Lie operator and the theorems are useful for many practical applications.

12. Lie Algebraic Factorization of Multivariable Evolution Operators:
Definition and the Solution of the Canonical Problem¹²

We have recently shown that the factorization of certain Lie algebraic evolution operators into a convergent infinite product of simple evolution operators is possible for one-dimensional cases. In this paper, we deal with the multivariable case. To this end, we formulate the factorization for the general case, then we show that most of the practical problems can be brought to a canonical one. The canonical problem has nothing different in concept but the relevant partial

differential equations to be solved can be easily handled. Two simple illustrative examples and the concluding remarks complete the work.

13. Global Sensitivity Analysis of Nonlinear Chemical Kinetic Equations
Using Lie Groups: I. Determination of One-Parameter Groups¹³

We introduce one-parameter groups of transformations that effect wide-ranging changes in the rate constants and input/output fluxes of homogeneous chemical reactions involving an arbitrary number of species in reactions of zero, first and second order. Each one-parameter group is required to convert every solution of such elementary rate equations into corresponding solutions of a one-parameter family of altered elementary rate equations. The generators of all allowed one-parameter groups are obtained for systems with N species using an algorithm which exactly determines their action on the rate constants, and either exactly determines or systematically approximates their action on the concentrations. Compounding the one-parameter groups yields all many-parameter groups of smooth time-independent transformations that interconvert elementary rate equations and their solutions.

14. Global Sensitivity Analysis of Nonlinear Chemical Kinetic Equations
Using Lie Groups: II. Some Chemical and Mathematical Properties of the
Transformation Groups¹⁴

This paper establishes a number of properties of transformation groups that map elementary kinetic equations into new elementary kinetic equations with altered rate constants. The chemical significance of the transformations is assessed by applying them to systems involving two

reacting species. There are then twelve one-parameter groups of mappings. Some mappings may be used to study the effects of changes in input/output fluxes on concentrations and their compensation by changes in other rate constants. A number of mappings transform nonlinear kinetics into approximately linear kinetics valid in regions larger than those obtained by standard methods. In some cases, the linearization is globally exact. Some mappings created lumped concentration variables and may be used to systematically reduce the number of manifest concentration variables in nonlinear, as well as linear, kinetic equations. The global mappings may be characterized by the functions of rate constants and functions of concentrations that they leave invariant. Although they produce large changes in rate constants and concentrations, none of these mappings change the topology of concentration phase plots as they map a phase plot determined by one set of initial conditions and rate constants into that determined by transformed initial conditions and rate constants. Metrical properties of the concentration maps generally depend upon the accuracy with which the group generators are approximated; systematic methods for their improvement are sketched.

IV. Participating Professional Personnel.

Research Staff: Dr. Richard Yetter and Dr. S-Y. Cho

Visiting Professional Collaborators:

Prof. Carl Wulfman, Prof. Metin Demiralp and Prof. Sandor Vajda

Postdoctoral Associate: Dr. Richard Hedges

Graduate Student: Mr. Genyuan Li

V. Presentations.

Prof. Rabitz gave invited presentations for the Workshop for Theoretical Chemistry, Utah; NASA Ames; Battelle Northwest Laboratories; ACS Boston; American Conference on Informational Sciences and Systems; Wright Patterson Air Force Base; and during an extensive trip to China for the International Symposium on Modern Chemistry (Fudan University, Shanghai-Jiao Tong University, Beijing University and the Institute of Chemistry, Beijing).

Dr. Yetter gave invited talks at the University of Kentucky; the Third International Workshop on Reduced Chemical Kinetic Mechanisms and Asymptotic Approximations, Cambridge University, Cambridge, England;, and at the Hazardous Substance Management Research Center, New Jersey Institute of Technology, Newark, NJ.

VI. Inventions

None

REFERENCES

1. S. Vajda, H. Rabitz, and R.A. Yetter, Comb. and Flame, 82, 270 (1990).
2. S. Vajda and H. Rabitz, Chem. Eng. Sci., submitted.
3. R. Yetter, H. Rabitz and R. Hedges, Int. J. of Chem. Kinetics, 23, 251 (1991).
4. M. Mishra, L. Peiperl, Y. Reuven, H. Rabitz, R. Yetter, and M. Smooke, J. Phys. Chem., in press.
5. M. Mishra, R. Yetter, Y. Reuven, H. Rabitz, and M. Smooke, Int. J. of Chem. Kinetics, submitted.
6. G. Li and H. Rabitz, Chem. Eng. Sci., 45, 977 (1990).
7. G. Li and H. Rabitz, Chem. Eng. Sci., 46, 95 (1990).
8. G. Li and H. Rabitz, Chem. Eng. Sci., 44, 1413 (1989).
9. G. Li and H. Rabitz, Chem. Eng. Sci., 46, 583 (1991).
10. G. Li and H. Rabitz, Chem. Eng. Sci., in press.
11. M. Demiralp and H. Rabitz, Int. J. Eng. Sci., in press.
12. M. Demiralp and H. Rabitz, Int. J. Eng. Sci., in press.
13. C.E. Wulfman and H. Rabitz, J. Math. Chem., 3, 243 (1989).
14. C.E. Wulfman and H. Rabitz, J. Math. Chem., 3, 261 (1989).

Appendix A

1. Effects of Thermal Coupling and Diffusion on the Mechanism of H_2 Oxidation in Steady Premixed Laminar Flames, S. Vajda, H. Rabitz, and R.A. Yetter, Comb. and Flame, 82, 270 (1990).

Effects of Thermal Coupling and Diffusion on the Mechanism of H_2 Oxidation in Steady Premixed Laminar Flames

S. VAJDA and H. RABITZ

Department of Chemistry, Princeton University, Princeton, NJ 08544

and

R. A. YETTER

Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544

The article considers the question why steady premixed laminar flames can be successfully described by highly reduced models, whereas the underlying mechanism is inherently complex. The calculations are performed on H_2 -air systems. Sensitivity functions are evaluated and studied for diffusion-free situations, both isothermal and adiabatic, as well as for steady premixed flames. In the diffusion-free cases most reactions of a 38-step mechanism are shown to be influential in a distinct fashion. The form of sensitivity functions is, however, radically changed and rendered self-similar by simultaneous thermal coupling and diffusion that introduce strong nonlinear coupling among the variables. Due to self-similarity, the mechanism can be reduced to 15 reactions, while keeping the temperature profile and the mass fraction profiles of molecular species almost unchanged in flame calculations. Furthermore, there exists an invariant subspace in the space of kinetic parameters such that large parameter perturbations along any vector in this subspace result in relatively small changes of the computed flame properties. By giving mechanistic interpretation to such parameter perturbations, the model can be simplified in many ways. In particular, a sequence of models is constructed in the stoichiometric H_2 -air flame problem that converge to a nine-step reduced mechanism with quasi-steady-state assumptions in radicals except H, thereby resulting in a two-step quasi-global model. All these approximations are unfeasible without the presence of molecular and thermal diffusion.

INTRODUCTION

It was a well-known truism among kineticists that "If one wishes to understand combustion reactions, one does not study combustion" [1]. A detailed understanding has been achieved for many combustion reactions (see, for example, Ref. 2), and now we face the problem of using this large amount of kinetic information when modeling a particular process with coupled kinetic, thermal, and diffusion phenomena. Most results of combustion science are based on the

assumption that the two latter processes will admit the use of simplified kinetic models [3]. In fact, the computational cost of a treatment involving a detailed mechanism would be too great in many multidimensional applications. In addition, the existence, multiplicity, stability, and structure of traveling-wave (steady-flame) solutions are difficult to explore solely via simulations, and the asymptotic-analytic treatment of highly reduced models with one or two global reactions has had an enormous impact on the understanding of these phenomena (see Ref. 4 and the contributions to

Refs. 5 and 6). Recent efforts have been devoted to systematic reduction of combustion mechanisms [7-9], offering procedures for constructing global stoichiometric and kinetic equations through the use of simplifying assumptions such as quasi-steady-state relations for certain intermediates and partial equilibrium of certain reactions.

Although emphasizing the success of simplified models in combustion, it is interesting to recall the somewhat contradictory status of the quasi-steady-state approximation (QSSA) in chemical kinetics. Though the QSSA has been the most important technique in elucidating reaction mechanisms since its formulation by Bodenstein, its validity and usefulness have also been questioned [10-14], and considerable efforts have been devoted to formulating conditions for its use (see, e.g., Refs. 15-20). It is easy to verify that many combustion reactions, with radical concentrations comparable to those of the reactants and products, do not pass these tests. Further factors that might invalidate the QSSA treatment are an overly short residence time in the flame for the radical concentrations to reach steady-state values, and the diffusion of radicals away from the regions of maximum radical concentrations [3, p. 129].

In spite of the above problems, excellent predictions have been reported in flame calculations involving the QSSA (e.g., [7, 8, 21, 22]). The analysis of this apparent contradiction is the main issue of the present article, considering the example of H_2 oxidation and generalizing the numerical results. Our first goal is to study the influence of heat release and diffusion on the relative importance of elementary reactions. The techniques involved are sensitivity analysis, now a routine tool for selecting the most influential part of a mechanism [7, 23, 24], and principal component analysis, which also reveals the applicable simplifying assumptions [25, 26].

The article is organized as follows. In section 2 we list the elementary reactions used here to describe H_2 oxidation under different conditions and write the governing equations. Section 3 is a summary of computational methods. To study the "pure" kinetic phenomena, in section 4 the isothermal, diffusion-free situation is considered.

The mechanism is shown to be inherently complex, i.e., most reactions of the starting mechanism are influential and should be retained. In section 5 we proceed to the adiabatic, diffusion-free system to determine the influence of thermal coupling on the relative importance of elementary reactions. Diffusion is first considered in section 6, where sensitivity functions for the steady, isobaric, quasi-one-dimensional, premixed laminar H_2 -air flame are computed and studied. Though the temperature is known to be a dominant variable in combustion processes, we show that only the simultaneous effects of thermal and transport phenomena change the form of the sensitivity function significantly, leading to their self-similarity. This interesting property [27] is exploited for mechanism reduction and for kinetic model simplification in sections 7 and 8, respectively. In particular, the concept of self-similarity enables us to explain the validity of simplifying assumptions in steady premixed flames that would be completely unfeasible in diffusion-free situations. Although numerical results are presented mostly for the stoichiometric H_2 -air flame, we try to draw more general conclusions by subsequent theoretical analysis.

Reaction Mechanism and Flame Model

The elementary reactions in the mechanism of H_2 oxidation have been extensively studied and documented. The special interest in this system is due to the fact that although the mechanism is much smaller than for hydrocarbon oxidation, the same reaction steps are also essential for the combustion of the latter ones. In addition, H_2 is itself a practical fuel, currently being considered to fuel the aerospace plane.

The mechanism is not discussed here because there exist a number of comprehensive reviews [2, 28]. The reactions listed in Table 1 as input data for the chemical kinetics interpreter of the CHEMKIN program [29] are based on Refs. 2 and 30, and they represent the influential subset of a much larger set of reactions that can occur theoretically [31]. For completeness we consider 19 pairs of forward/backward reactions, although

TABLE 1
Reaction Mechanism and Arrhenius Parameters for Hydrogen Oxidation

No.	Reaction ^{a,b}	A^c	n	E
1.	$H + O_2 \rightarrow O + OH$	1.64(14)	0	15470.
2.	$O + OH \rightarrow H + O_2$	0.89(11)	0.387	-1689.
3.	$O + H_2 \rightarrow H + OH$	5.08(4)	2.67	6292.
4.	$H + OH \rightarrow O + H_2$	2.88(4)	2.64	4473.9
5.	$H_2 + OH \rightarrow H_2O + H$	6.30(6)	2.00	2961.
6.	$H_2O + H \rightarrow H_2 + OH$	6.77(7)	1.89	18291.3
7.	$O + H_2O \rightarrow OH + OH$	3.98(9)	1.32	16150.8
8.	$OH + OH \rightarrow O + H_2O$	2.10(8)	1.40	-397.4
9.	$H + H + M \rightarrow H_2 + M$	1.08(20)	-1.67	822.7
10.	$H_2 + M \rightarrow H + H + M$	4.58(19)	-1.4	104400.
11.	$O + O + M \rightarrow O_2 + M$	6.17(15)	-0.5	0.
12.	$O_2 + M \rightarrow O + O + M$	4.94(17)	-0.65	118909.
13.	$O + H + M \rightarrow OH + M$	4.72(18)	-1.0	0.
14.	$OH + M \rightarrow O + H + M$	1.13(18)	-0.76	101751.
15.	$H + OH + M \rightarrow H_2O + M$	2.25(22)	-2.0	0.
16.	$H_2O + M \rightarrow H + OH + M$	1.02(23)	-1.84	118899.
17.	$H + O_2 + M \rightarrow HO_2 + M$	2.00(15)	0.	-1000.
18.	$HO_2 + M \rightarrow H + O_2 + M$	4.47(15)	-0.074	50388.9
19.	$H + HO_2 \rightarrow H_2 + O_2$	6.63(13)	0.	2126.
20.	$H_2 + O_2 \rightarrow H + HO_2$	1.25(13)	0.35	54305.7
21.	$H + HO_2 \rightarrow OH + OH$	1.69(14)	0.	874.
22.	$OH + OH \rightarrow H + HO_2$	5.39(10)	0.71	34078.4
23.	$HO_2 + OH \rightarrow H_2O + O_2$	1.45(16)	-1.	0.
24.	$H_2O + O_2 \rightarrow HO_2 + OH$	2.94(16)	-0.76	67510.4
25.	$HO_2 + O \rightarrow O_2 + OH$	1.81(13)	0.	-397.
26.	$O_2 + OH \rightarrow HO_2 + O$	1.93(12)	0.32	49965.3
27.	$HO_2 + HO_2 \rightarrow H_2O_2 + O_2$	1.00(13)	0.	1000.
28.	$H_2O_2 + O_2 \rightarrow HO_2 + HO_2$	1.22(15)	-0.36	34715.7
29.	$H_2O_2 + OH \rightarrow H_2O + HO_2$	7.00(12)	0.	1430.
30.	$H_2O + HO_2 \rightarrow H_2O_2 + OH$	1.16(11)	0.6	35224.5
31.	$H_2O_2 + H \rightarrow H_2O + OH$	1.00(13)	0	3590.
32.	$H_2O + OH \rightarrow H_2O_2 + H$	5.30(7)	1.31	70588.8
33.	$H_2O_2 + H \rightarrow HO_2 + H_2$	4.82(13)	0.	7948.
34.	$HO_2 + H_2 \rightarrow H_2O_2 + H$	7.45(10)	0.71	26411.4
35.	$H_2O_2 + M \rightarrow OH + OH + M$	1.20(17)	0.	45500.
36.	$OH + OH + M \rightarrow H_2O_2 + M$	1.40(11)	1.15	-6403.9
37.	$O + OH + M \rightarrow HO_2 + M$	1.00(17)	0.	0.
38.	$HO_2 + M \rightarrow O + OH + M$	7.49(19)	-0.47	68546.6

^a $[M] = [N_2] + [O_2] + 16[H_2O] + 2.5[H_2] + [HO_2] + [H_2O_2] + [H] + [O] + [OH]$.

^b Units are centimeters, moles, seconds, and calories.

^c Numbers in parentheses denote powers of ten.

one of the reactions in certain pairs is negligible under all conditions studied in this article and are omitted as part of the mechanism reduction process (see below). As detailed in Ref. 30, in each pair we choose the rate of that reaction (forward or backward) for which more reliable data are available, whereas the other rate constant is cal-

culated from the equilibrium data of the JANAF Thermochemical Tables [32]. The rate coefficients follow the modified Arrhenius temperature dependence

$$k_j = A_j T^{n_j} e^{-E_j/RT}, \quad (1)$$

with the parameters A_j , n_j , and E_j listed in

Table 1 being consistent with the equilibrium data.

Our formulation of the premixed flame problem closely follows the one given by Smooke [33-35]. Upon neglecting viscous effects, body forces, radiative heat transfer, and the diffusion of heat due to the concentration gradients, the equations governing steady, isobaric, one-dimensional flame propagation are

$$\dot{M} = \rho u = \text{const}, \quad (2)$$

$$\dot{M} \frac{dY_k}{dx} = - \frac{d}{dx} (\rho Y_k V_k) + \dot{\omega}_k W_k, \quad (3)$$

$$k = 1, 2, \dots, K,$$

$$\dot{M} \frac{dT}{dx} = \frac{1}{c_p} \frac{d}{dx} \left(\lambda \frac{dT}{dx} \right) - \frac{1}{c_p} \sum_{k=1}^K \rho Y_k V_k c_{p,k} \frac{dT}{dx} - \frac{1}{c_p} \sum_{k=1}^K \dot{\omega}_k h_k W_k, \quad (4)$$

coupled with the equation of state

$$\rho = \frac{P \bar{W}}{RT}. \quad (5)$$

In these equations x denotes the independent spatial coordinate, \dot{M} is the (constant) mass flow rate, T is the temperature, Y_k is the mass fraction of the k th species; P is the pressure, u is the velocity of the mixture, ρ is the mass density, W_k is the molecular weight of the k th species, \bar{W} is the mean molecular weight of the mixture; λ is the thermal conductivity, $c_{p,k}$ is the specific heat of the k th species at constant pressure, $\dot{\omega}_k$ is the molar rate of production of the k th species per unit volume, h_k is the specific enthalpy of the k th species, and V_k is the diffusion velocity of the k th species, approximated with a Fickian relationship [34]. As in Refs. 34 and 35, the flame problem is posed on the finite interval $0 \leq x \leq L$. The boundary conditions are given by

$$T(0) = T_b, Y_k(0) = \epsilon_k(0), k = 1, 2, \dots, K \quad (6)$$

and

$$\frac{dT}{dx}(L) = 0, \frac{dY_k}{dx}(L) = 0, k = 1, 2, \dots, K, \quad (7)$$

where T_b is the temperature of the unburned gas, and the known mass flux fraction of the k th species is defined as

$$\epsilon_k = Y_k + \frac{\rho Y_k V_k}{\dot{M}} \quad (8)$$

As suggested by Smooke [34, 35], the mass flow rate \dot{M} in a freely propagating (adiabatic) flame is determined by introducing the additional differential equation

$$\frac{d\dot{M}}{dx} = 0, \quad (9)$$

and the boundary condition

$$T(x_f) = T_f, \quad (10)$$

where x_f is a specified spatial coordinate $0 < x_f < L$, and T_f is a specified temperature. To study the diffusion-free system we set $V_k = 0$ and $\lambda = 0$ in Eqs. 3 and 4. In this case the mass flow rate \dot{M} is assigned, and we obtain an initial value problem with the initial conditions in Eq. 6, where $\epsilon_k = Y_k$ according to Eq. 8. In calculations with a fixed temperature profile (e.g., in the isothermal case) only Eqs. 3 are considered.

Methods of Analysis

The initial value problems are solved by a semi-implicit Runge-Kutta method [36]. The normalized sensitivity coefficients $\partial \ln Y_k / \partial \ln A_j$ and $\partial \ln T / \partial \ln A_j$ are computed with a decomposed direct method [37] in conjunction with the same ODE-solver. The required derivatives are generated by subroutines of the CHEMKIN package [29].

The solution of the flame problem involves a finite difference approximation of the derivatives in Eqs. 3 and 4 on an adaptively determined computational mesh [33-35]. As in Ref. 35, we determine the value of \dot{M} for the adiabatic flame

problem and compute the sensitivity coefficients $\partial \ln u / \partial \ln A_j$. Then M is fixed at the obtained value, and the sensitivity coefficients $\partial \ln Y_k / \partial \ln A_j$ and $\partial \ln T / \partial \ln A_j$ are computed. In addition to a slightly different reaction mechanism, the only deviation from Ref. 35 is that independent parameter perturbations are considered in the sensitivity analysis. Therefore, we obtain a sensitivity coefficient for each elementary reaction separately, instead of the total sensitivity coefficients used in Refs. 35 and 38. Having separate sensitivity functions is of considerable importance for the purposes of this article, particularly for the analysis of simplifying assumptions.

With 10 variables (9 species plus the temperature), 38 rate coefficients, and 10 further parameters (the thermal conductivity λ and 9 diffusion coefficients), each sensitivity calculation results in 480 sensitivity functions, in addition to the 48 flame speed sensitivity coefficients. It is a formidable task to analyze such a large amount of numerical information. In addition, we show that the simple inspection of the sensitivity functions may be somewhat misleading. Principal component analysis [25] offers a compact way of exhibiting the kinetic information hidden in sensitivity results. The method is based on introducing a response function of the form

$$Q(\mathbf{p}) = \sum_{j=1}^q \sum_{i=1}^m \left[\frac{Y_i(x_j, \mathbf{p}) - Y_i(x_j, \mathbf{p}^0)}{Y_i(x_j, \mathbf{p}^0)} \right]^2, \quad (11)$$

where $Y_i(x_j, \mathbf{p})$ denotes the i th variable (i.e., mass fraction or temperature) of interest at the mesh point x_j and parameter value \mathbf{p} , with \mathbf{p}^0 being the nominal parameter value where the analysis is carried out. In Eq. 11, q and m , respectively, denote the number of mesh points and the number of variables considered in the principal component analysis. The function $Q(\mathbf{p})$ is then a measure of the total change in the variables Y_1, \dots, Y_m brought about by the variation $\Delta \mathbf{p} = \mathbf{p} - \mathbf{p}^0$ in the parameters. Let \mathbf{S} denote the $(m \times q) \times r$ matrix of the normalized sensitivity coefficients $\partial \ln y_i(x_j, \mathbf{p}^0) / \partial \ln$

p_l , where $i = 1, \dots, m$, $j = 1, \dots, q$, $l = 1, \dots, r$, and r denotes the number of the parameters. Then the Gauss approximation, well known in nonlinear least squares parameter estimation, yields

$$Q(\alpha) \approx \tilde{Q}(\alpha) = (\Delta \alpha)^T \mathbf{S}^T \mathbf{S} (\Delta \alpha), \quad (12)$$

where $\alpha_j = \ln p_j$, $\alpha_j^0 = \ln p_j^0$, and $\Delta \alpha = \alpha - \alpha^0$ (see [25] for details). Now we introduce the new coordinates $\psi = \mathbf{U}^T \alpha$ in the space of logarithmic parameters, where \mathbf{U} denotes the matrix of the normalized eigenvectors of $\mathbf{S}^T \mathbf{S}$, and the ψ s are called principal components. In terms of these new coordinates the quadratic function in Eq. 12 is transformed to the normal form

$$\tilde{Q}(\alpha) = \sum_{i=1}^r \lambda_i (\Delta \psi_i)^2, \quad (13)$$

where $\Delta \psi = \mathbf{U}^T \Delta \alpha$, and $\lambda_1 > \lambda_2 > \dots > \lambda_r$ are the eigenvalues of the matrix $\mathbf{S}^T \mathbf{S}$. Equation 13 gives a decomposition of the space of logarithmic parameters α into "influential" and "noninfluential" subspaces. If we make a step of unit length from the point α^0 along an eigenvector \mathbf{u}_i , i.e., $\Delta \psi_i = 1$, then $\tilde{Q}(\alpha)$, and hence λ_i measures the significance of reactions that are present in the principal component ψ_i . Principal components corresponding to the large eigenvalues define the influential part of the mechanism.

Important mechanistic interpretation can be given to certain forms of eigenvectors. For example, assume that the normalized eigenvector \mathbf{u}_1 corresponding to the largest eigenvalue λ_1 is given by $\mathbf{u}_1 = (0.707, -0.707, 0, \dots, 0)^T$. To move along \mathbf{u}_1 we select $\Delta \alpha_2 = -\Delta \alpha_1$, while the other parameters are unperturbed. This implies $\ln p_2 + \ln p_1 = \ln p_2^0 + \ln p_1^0$, and hence moving along the curve $p_1 p_2 = \text{const}$ in the space of original parameters. Thus, the largest change in the response function Q is attained by increasing one of the parameters, say p_1 , while decreasing p_2 in order to keep their product constant. This will be the typical situation we find with p_1 and p_2 denoting the rate constants of competing reactions.

Important conclusions can be drawn also from the existence of the eigenvector $\mathbf{u}_i = (0.707,$

0.707, 0, ..., 0)^T corresponding to a small eigenvalue $\lambda_i \approx 0$. The line $\Delta\alpha_1 = \Delta\alpha_2$ defines the curve $p_1/p_2 = \text{const}$ in the space of the original parameters. Since $\lambda_i \approx 0$, we have $\tilde{Q}(\alpha) \approx 0$ along this curve. Thus, the response function (Eq. 11) depends only on the ratio p_1/p_2 and does not depend on p_1 and p_2 separately. If p_1 and p_2 are the rate coefficients of a forward/backward reaction pair, then this clearly indicates the validity of partial equilibrium assumption. Uncovering the dependencies among the parameters, principal component analysis proved to be very useful for uncovering, confirming, or denying the validity of simplifying approximations [25, 26].

As discussed in Ref. 25, an eigenvalue λ_i is classified as "small" if $\lambda_i < 10^{-4} mq$. Though this is an approximate rule, reactions that are present only in principal components corresponding to small eigenvalues usually have little influence on the behavior of the system.

Isothermal, Diffusion-free Conditions

The kinetics of stoichiometric H_2 -air system has been studied at $P = 1$ atm and $T = 920$ K,

slightly above the explosion limit. The selected mass flow rate is $\dot{M} = 0.175 \text{ g cm}^{-2} \text{ s}^{-1}$, which gives the flow speed $u = 631 \text{ cm s}^{-1}$ at the cold boundary. The mass fraction profiles are shown in Fig. 1. The normalized sensitivity coefficients $\partial \ln Y_i / \partial \ln A_j$ have been computed at $q = 30$ equidistant mesh points for $L = 10$ cm, and all species except N_2 are considered in the principal component analysis. Thus, $m = 8$ in Eq. 11, and the threshold value for "small" eigenvalues is $\lambda_{\min} = 2.4 \times 10^{-2}$. The eigenvalues exceeding this limit and the corresponding principal components are listed in Table 2. As discussed in section 3, the form of the eigenvector u_i corresponding to the very large eigenvalue λ_1 clearly shows that the most important part of the mechanism is the competition of reactions 1 and 17. The corresponding sensitivity functions of H_2 are so large that we had to plot them separately from the others in Fig. 2. The further most important reactions are 21, 19, 3, 5, 20, 27, 35, 31, 2, 15, and 37, which appear in the principal components ψ_2 - ψ_{10} . The sensitivity functions of H_2 with respect to the rate constants of these reactions are shown in Fig. 3.

The 16 influential principal components in

TABLE 2

Principal Components for Stoichiometric Mixture at $T = 920$ K, Mole Fractions for All Species

No.	Eigenvalue ^a	Parameters in the Principal Component ^b
1	2.64(+7)	1[0.72], 17[-0.69]
2	3.92(+3)	1[-0.30], 17[-0.30], 19[-0.46], 21[0.77]
3	1.85(+2)	1[0.54], 3[0.24], 17[0.58], 19[-0.42], 21[0.21]
4	1.19(+2)	3[0.51], 5[0.52], 19[0.29], 27[-0.25], 35[-0.28]
5	5.51(+1)	3[-0.52], 5[0.79]
6	3.40(+1)	1[0.24], 3[-0.60], 5[-0.25], 17[0.25], 19[0.31], 20[-0.25]
7	3.29(+1)	20[0.88], 35[-0.37]
8	2.21(+1)	19[0.20], 20[0.29], 27[-0.34], 35[0.80]
9	9.90(+0)	19[0.49], 21[0.39], 27[0.71]
10	7.11(+0)	2[0.66], 15[0.54], 31[-0.28], 37[0.22]
11	4.37(+0)	2[0.27], 27[0.31], 31[0.55], 33[0.24], 36[0.77]
12	3.17(+0)	34[0.95]
13	4.16(-1)	29[0.21], 31[0.51], 33[0.23], 36[0.77]
14	2.04(-1)	2[-0.48], 8[0.35], 15[0.54], 25[-0.55]
15	7.92(-2)	2[-0.42], 8[-0.23], 15[0.46], 25[0.55], 37[0.41]
16	2.34(-2)	7[0.23], 8[0.22], 23[-0.20], 25[0.26], 29[0.77], 31[-0.37], 37[-0.23]

^a Numbers in parentheses denote powers of ten.

^b Numbers in brackets denote the coefficients of the parameters in the corresponding principal component.

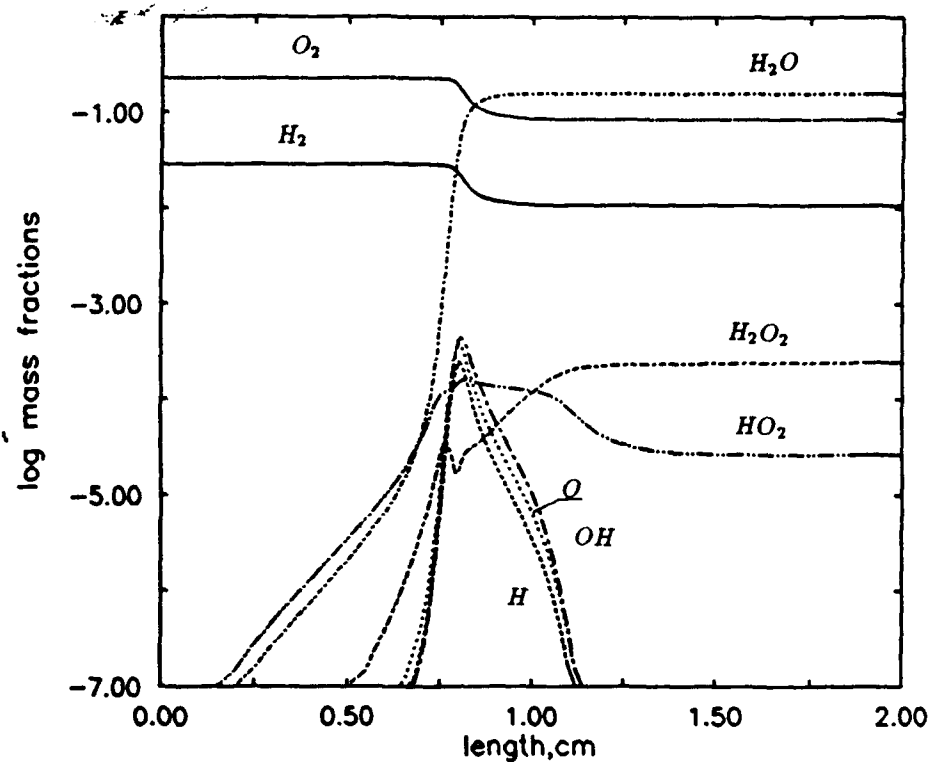


Fig. 1. Mass fractions in the isothermal system at $T = 920$ K.

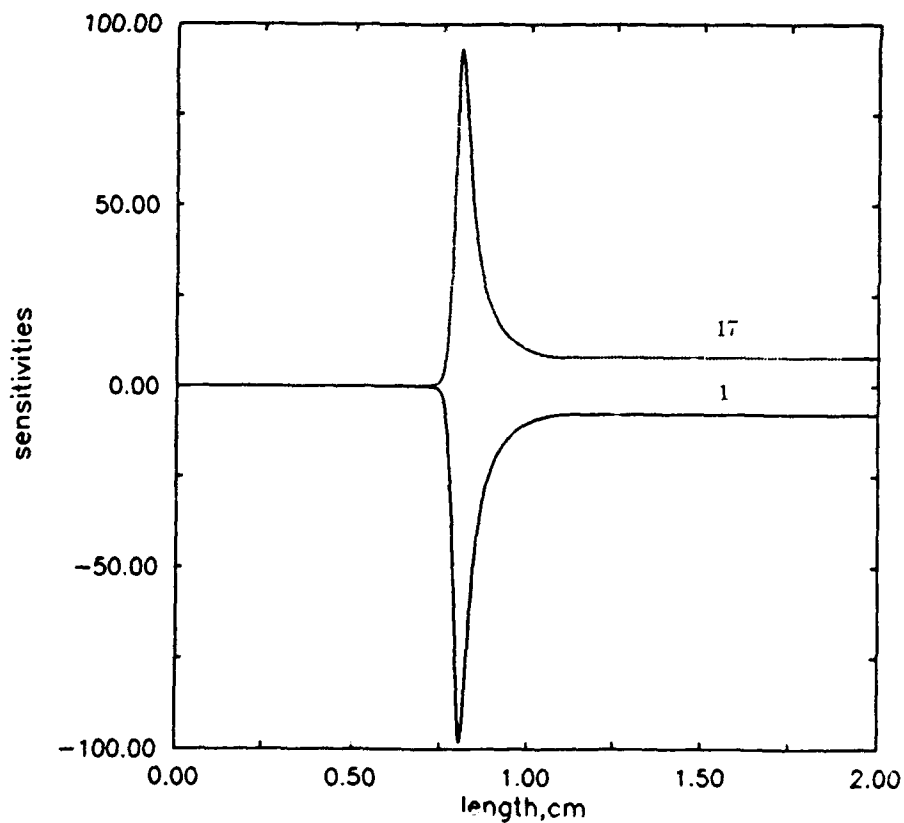


Fig. 2. Normalized sensitivity functions of the H_2 mass fraction for reactions 1 and 17 in the isothermal system.

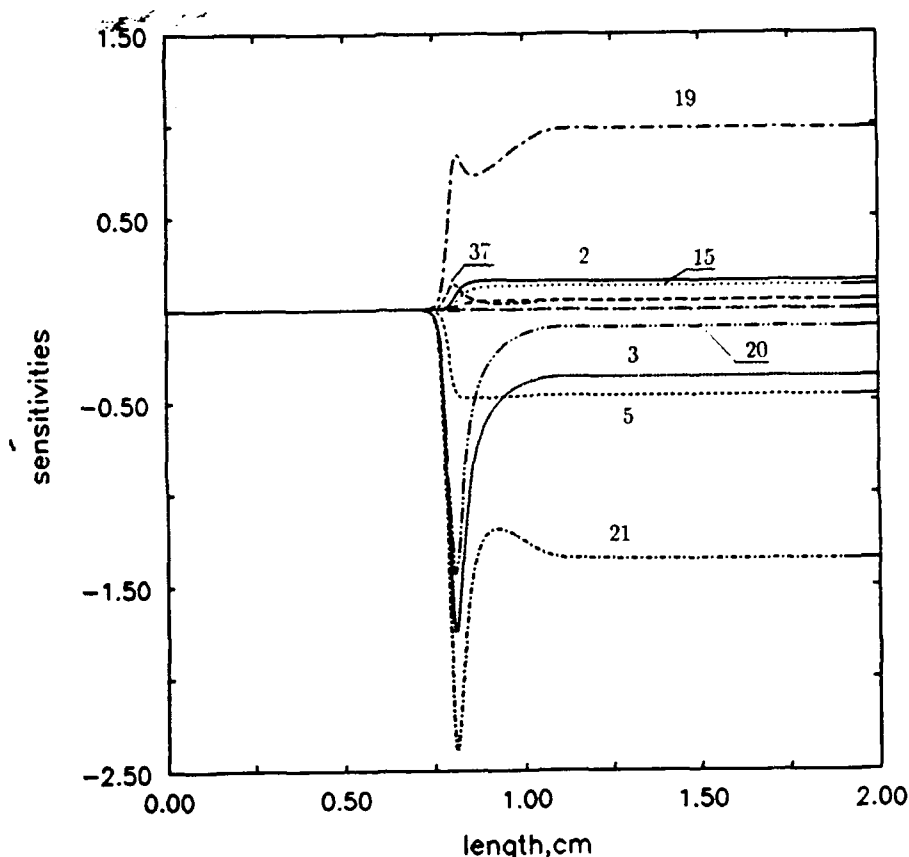


Fig. 3. Normalized sensitivity functions of the H_2 mass fraction for the further most important reactions in the isothermal system.

Table 2 contain only 21 steps {1, 2, 3, 5, 7, 8, 15, 17, 19, 20, 21, 23, 25, 27, 29, 31, 33, 34, 35, 36, 37} of the 38 in the starting mechanism. The mechanism of the 21 reactions gives rise to solutions that deviate less than 5% from the solutions of the complete model for all species (including the radicals) at all points of the considered interval [0, 10] cm.

We would like to further reduce the mechanism, even with the price of large errors in the radical concentrations. It is expected that further dispensable reactions can be found by restricting consideration to the sensitivity functions of those species whose behavior is to be preserved. In some cases this approach is very successful. For example, Edelson and Allara [39] ranked the 98 reactions of a low-temperature propane pyrolysis mechanism according to the absolute values of the sensitivity coefficients, computed only for a few species considered as experimental observables. It can be verified that the 52 reactions with

nonzero ratings give an excellent approximation for the concentrations of these "observable" species. In our previous article [25] it was, however, shown that considering only certain species can lead to erroneous conclusions in sensitivity analysis. Thus the approach has no general validity but is still worth a try. Therefore, we computed the principal components restricting consideration to the species H_2 , O_2 , and H_2O , considered observables in this work. As expected, a number of further reactions appears to be dispensable. Solving the kinetic differential equations we learned, however, that any further reduction of the 21-step mechanism results in large concentration deviations not only for the radicals but also for the "observable" species. For example, steps 7 and 23 are slightly influential according to Table 2 (they appear only in the principal component ψ_{16}) and seem to be dispensable when considering only the sensitivities of the "observables." Nevertheless, their elimination gives more

than 10% errors in the concentrations of the latter species. Because we want to find the simplest mechanism possible, this result is disappointing but easy to explain. Although the propane pyrolysis mechanism in Ref. 39 consists of several weakly connected subsystems (i.e., formation and removal of certain species that practically do not interact with the observable ones and only slightly influence the concentrations of the important radicals), all species in our starting H_2 oxidation mechanism are strongly coupled through the radical pool. It is exactly this strong coupling among all species that will enable us to simplify the mechanism in the presence of diffusion, as we show later in this article.

At this point, however, we have to conclude that the mechanism of H_2 oxidation under well-stirred isothermal conditions is inherently complex. The small eigenvalues are not listed in Table 2, because they do not reveal any dependencies among the retained 21 rate coefficients, and hence we must also exclude the validity of simplifying kinetic approximations such as the QSSA.

The minimal mechanism becomes even more complex if we want to extend its validity also to higher temperatures. Calculating the sensitivities at $T = 1500$ K and performing the principal component analysis shows that the set of influential reactions is {1-8, 15, 17, 19-21, 31, 33-37}. Although steps 17 and 19 are much less important than at $T = 920$ K, and several reactions consuming HO_2 lost their significance completely, the importance of the backward reactions 6 and 8 increases. Thus we must add these two steps to the minimal mechanism, consisting now of 23 reactions.

Adiabatic, Diffusion-free Conditions

The adiabatic calculations have been performed at $P = 1$ atm and $T_b = 920$ K at the cold boundary. The temperature and mass fraction profiles are shown in Fig. 4. The reaction is confined to a very narrow region. The prereaction and postreaction regions are almost isothermal, and the mass fraction profiles are similar to the ones

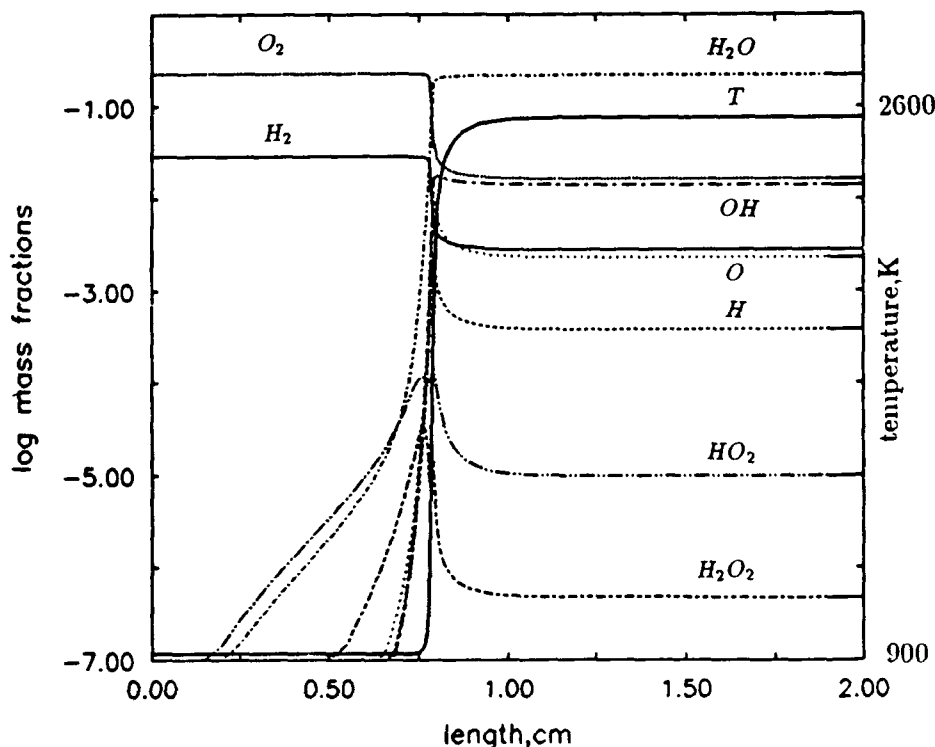


Fig. 4. Mass fractions and temperature in the adiabatic, diffusion-free system.

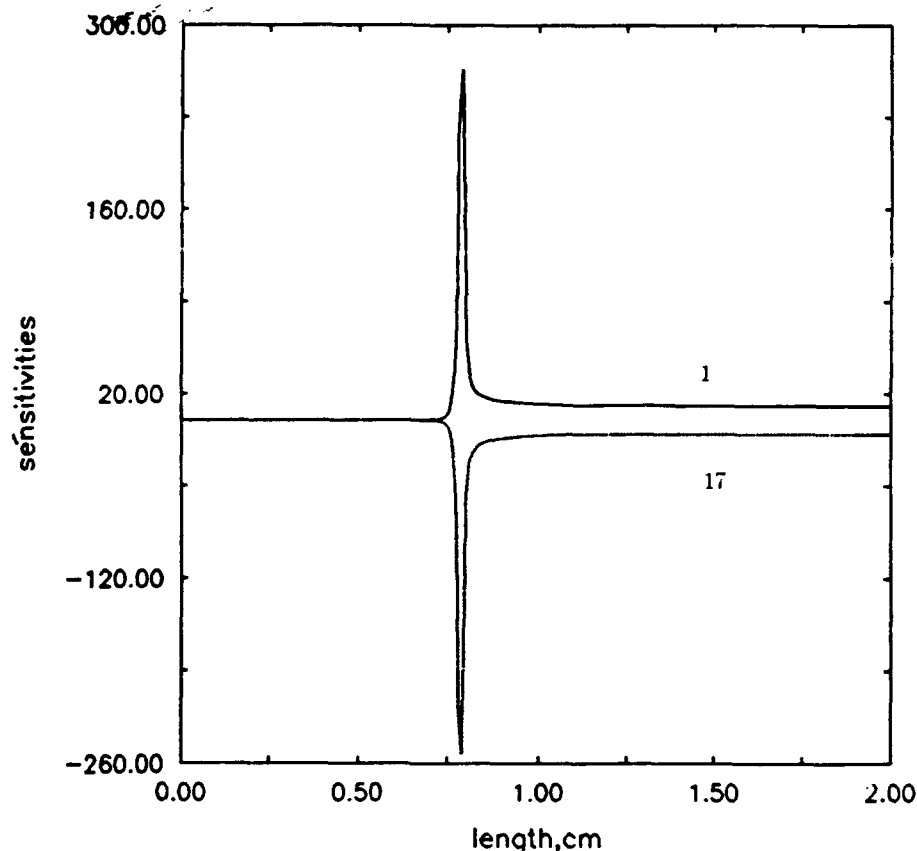


Fig. 5. Sensitivity functions of the temperature for reactions 1 and 17 in the adiabatic system.

found in the isothermal case for low and high temperatures, respectively.

As shown in Figs. 5 and 6, the temperature is very sensitive to parameter variations within the reaction zone. The nonvanishing "tail" of sensitivity functions indicates the influence of the parameters on the adiabatic temperature via a change in equilibrium. According to Fig. 5, the dominant part of the mechanism is again the competition of steps 1 and 17 for H^{\cdot} , giving rise to a very large eigenvalue in the principal component analysis.

Because the temperature is an important variable, we expected that the mechanism could be somewhat reduced by eliminating some reactions that do not significantly contribute to the heat release. This expectation, however, failed. As confirmed by the outcome of principal component analysis, any reaction important in isothermal oxidation either at low or at high temperature is also important in the adiabatic process. Thus we need the 23-step minimal mechanism derived in

the previous section. This mechanism can neither be reduced nor simplified through the use of kinetic approximations if we want to reproduce the "observable" variables (i.e., the temperature and the mass fractions for H_2 , O_2 , and H_2O) within 5% errors.

We introduce a decomposition of the sensitivity functions that shows the role of the temperature in the adiabatic system and explains why the mechanism cannot be reduced. For the sake of notational simplicity write Eqs. 3 and 4 in the diffusion-free case as

$$\begin{aligned} \frac{dY_k}{dx} &= f_k(Y_1, \dots, Y_K, T, p), \\ Y_k(0) &= Y_{k,b}, \quad k = 1, \dots, K, \end{aligned} \quad (14)$$

and

$$\frac{dT}{dx} = f_T(Y_1, \dots, Y_K, T, p), \quad T(0) = T_b, \quad (15)$$

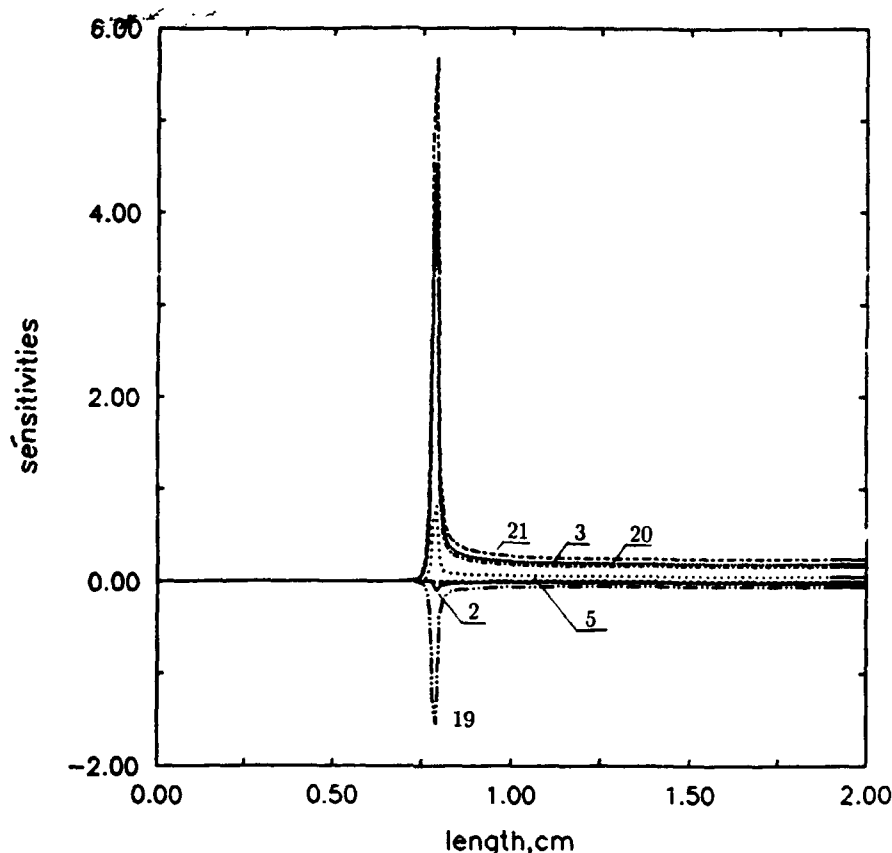


Fig. 6. Sensitivity functions of the temperature for the further most important reactions in the adiabatic system.

respectively. It will be convenient to write the K equations in Eq. 14 also in the vector form

$$\frac{d\mathbf{Y}}{dx} = \mathbf{f}(\mathbf{Y}, T, \mathbf{p}), \quad (16)$$

where $\mathbf{Y} = (Y_1, \dots, Y_K)^T$ and $\mathbf{f} \approx (f_1, \dots, f_K)^T$. Differentiating Eq. 16 with respect to the parameter p_j we obtain the sensitivity equation

$$\begin{aligned} \frac{d}{dx} \frac{\partial \mathbf{Y}}{\partial p_j} &= \frac{\partial \mathbf{f}}{\partial \mathbf{Y}} \frac{\partial \mathbf{Y}}{\partial p_j} \\ &+ \frac{\partial \mathbf{f}}{\partial T} \frac{\partial T}{\partial p_j} + \frac{\partial \mathbf{f}}{\partial p_j}, \end{aligned} \quad (17)$$

where $\partial \mathbf{f} / \partial \mathbf{Y}$ is the $K \times K$ Jacobian matrix of Eq. 16. As is well known, the sensitivity equation (Eq. 17) can be solved through the Green's function matrix $\mathbf{G}_1(x, x')$, which is the solution of

the matrix differential equation

$$\begin{aligned} \frac{d}{dx} \mathbf{G}_1(x, x') &= \frac{\partial \mathbf{f}}{\partial \mathbf{Y}}(x) \mathbf{G}_1(x, x') \\ &+ \mathbf{I} \delta(x - x'), \\ \mathbf{G}_1(x', x') &= 0, \end{aligned} \quad (18)$$

where \mathbf{I} is the $K \times K$ unit matrix and δ denotes the Dirac impulse function (see, e.g., Refs. 40 and 41). In terms of $\mathbf{G}_1(x, x')$ the sensitivity functions are given by

$$\begin{aligned} \frac{\partial \mathbf{Y}}{\partial p_j}(x) &= \int_0^x \mathbf{G}_1(x, x') \frac{\partial \mathbf{f}}{\partial T}(x') \frac{\partial T}{\partial p_j}(x') dx' \\ &+ \int_0^x \mathbf{G}_1(x, x') \frac{\partial \mathbf{f}}{\partial p_j}(x') dx'. \end{aligned} \quad (19)$$

Let us now fix the temperature at its adiabatic

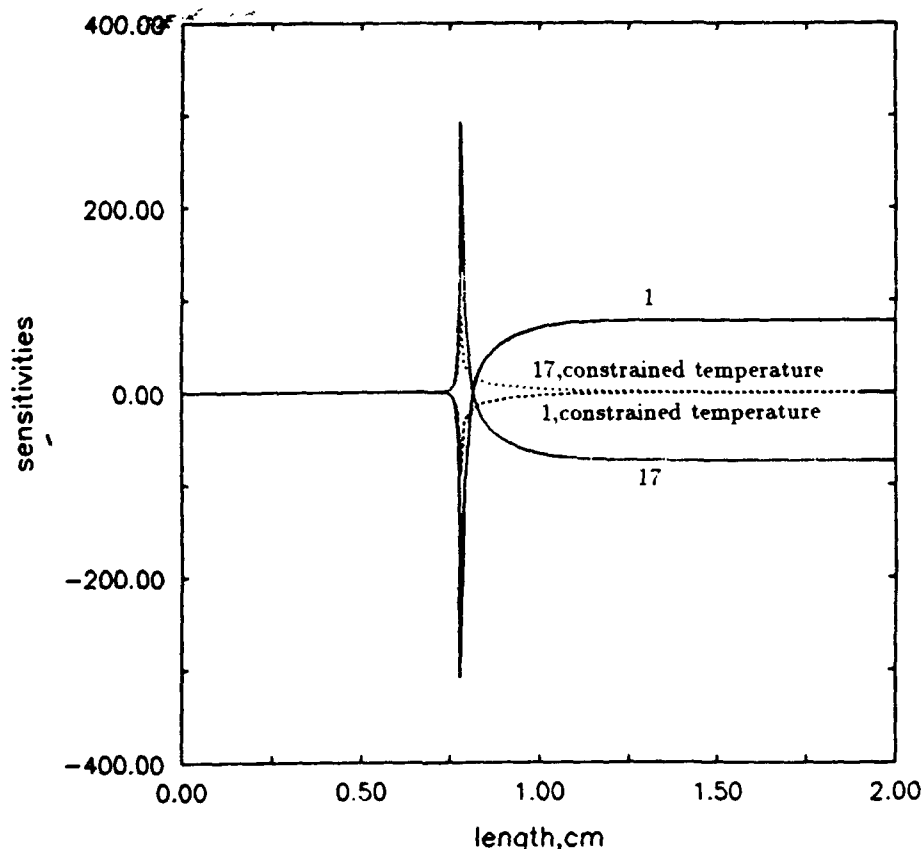


Fig. 7. Adiabatic (solid line) and constrained temperature (dashed line) sensitivity functions of the H_2 mass fraction for reactions 1 and 17 in the adiabatic system.

profile, i.e., consider the temperature as an external variable. Then $\partial T / \partial p_j = 0$, and the sensitivity functions are reduced to the second term on the right-hand side of Eq. 19. Thus this term, called the constrained temperature sensitivity function, has a well-defined physical meaning. Figure 7 shows the adiabatic sensitivity function (solid curves) and the constrained temperature ones (dashed curves) of the H_2 mass fraction for the most important steps 1 and 17. It follows from the temperature profile in Fig. 4 that the constrained temperature sensitivity functions are essentially isothermal ones, for low and high temperature in the prereaction and postreaction regions, respectively. Indeed, the dashed curves in Fig. 7 are similar to the ones in Fig. 2, both in form and magnitude. The only deviation is that the constrained temperature sensitivity functions of steps 1 and 17 vanish in the postreaction zone due to the high temperature, as discussed in the previous section.

While the constrained temperature sensitivity

coefficients measure the direct, quasi-isothermal effects of parameter perturbations, the first term in Eq. 19 corresponds to the indirect effects (i.e., the parameter perturbations that change the temperature profile, which, in turn, affects the mass fractions through the reactions). According to Fig. 7 this indirect influence of the temperature is more important than the direct one in the reaction zone. Equation 19 also explains the form of adiabatic sensitivity functions. It may be readily verified that for H_2 (the first component in the Y vector) the function $\partial f_1 / \partial T$ is positive in the prereaction zone and negative in the postreaction one. As shown in Fig. 5, $\partial T / \partial A_1 \neq 0$ and $\partial T / \partial A_{17} \neq 0$ in the postreaction zone, and the first integral in Eq. 19 results in the marked "overshoot" of the adiabatic sensitivity functions in Fig. 7.

Similar results are found for the other reactions. Figures 8 and 9 show the adiabatic and constrained temperature sensitivity functions, respectively, of H_2 for the second group of most

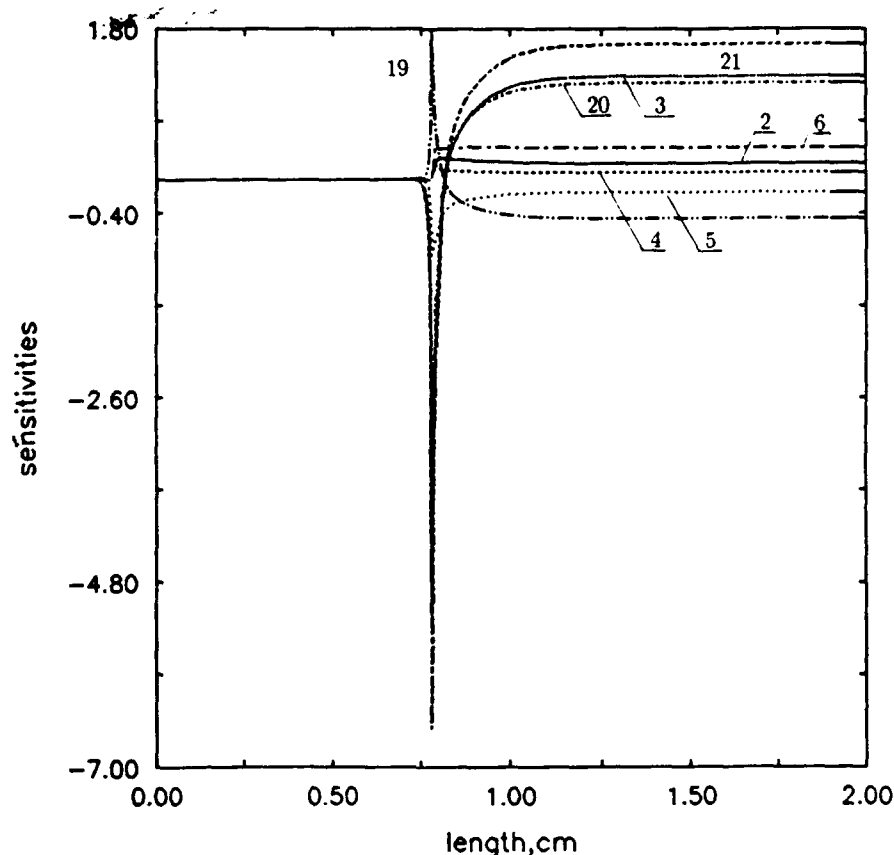


Fig. 8. Adiabatic sensitivity functions of the H_2 mass fraction for the reactions considered in Fig. 6.

important reactions 21, 19, 20, 3, 5, 6, 9, and 2, revealed by principal component analysis. The constrained sensitivities in Fig. 9 are similar to the isothermal ones in Fig. 3. It is easy to explain the deviations: steps 21 and 19 lose their significance, since at high temperature much less HO_2 is produced by reaction 17, whereas the backward reactions 4 and 6 become more important, as discussed in the previous section. The "overshoot" of the adiabatic sensitivity functions 21, 3, and 20 in Fig. 8 follows from the nonvanishing "tail" of the temperature sensitivity functions for these reactions, shown in Fig. 6. On the other hand, if the temperature sensitivity is small (e.g., for step 2), then the adiabatic and the constrained temperature sensitivity functions in Figs. 8 and 9, respectively, are identical, although this is somewhat masked by the different scales on the two plots.

The most important fact we can learn from the decomposition (Eq. 19) is the relative magnitude of the two terms. Although the indirect effect of

parameter perturbations is larger by a factor of 3 than the direct, quasi-isothermal pathway, this latter is definitely not negligible. Thus the adiabatic system retains all the complexities of the mechanism that is valid for both low and high temperatures under isothermal conditions. This observation gets added importance in comparing to the flame problem, where we encounter a completely different ratio of the two terms. We note that the temperature sensitivity functions shown in Fig. 6 are similar. The similarity is, however, weaker among the H_2 sensitivity functions in Fig. 8, and even such weak similarity was not observed among the sensitivities of further species, not shown here.

Steady Premixed Laminar Flame

Figure 10 shows the solutions of Eqs. 2-5 for the stoichiometric H_2 -air flame at $P = 1$ atm and cold boundary condition $T_b = 298$ K. The additional boundary condition (Eq. 10) is given by

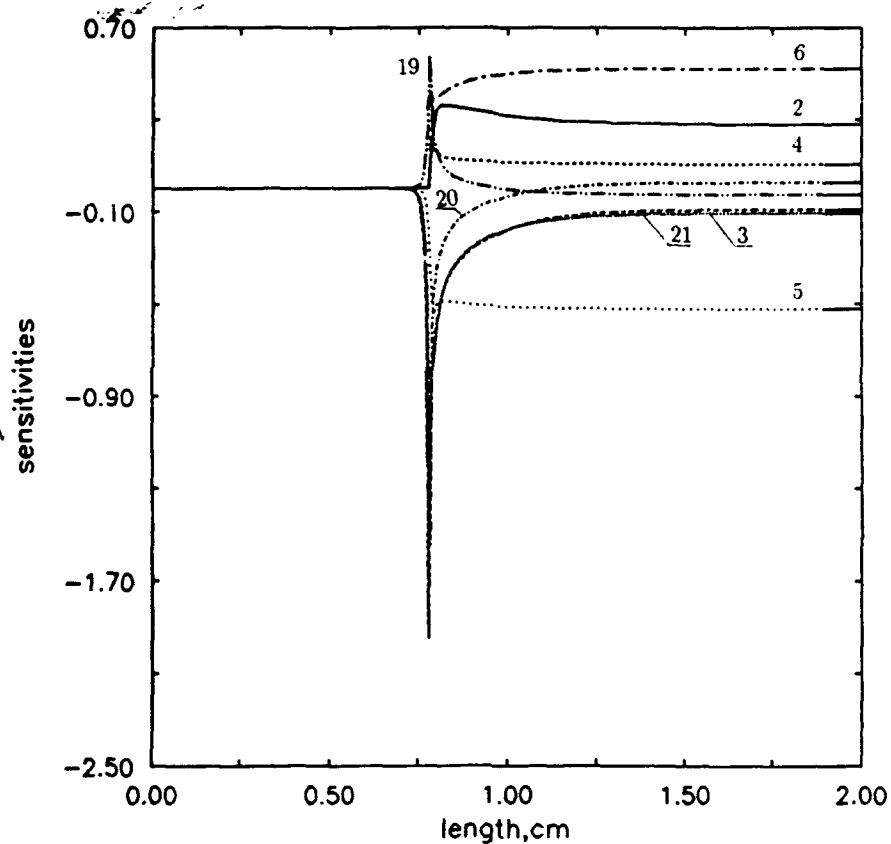


Fig. 9. Constrained temperature sensitivity functions of the H_2 mass fraction for the reactions considered in Figs. 6 and 8.

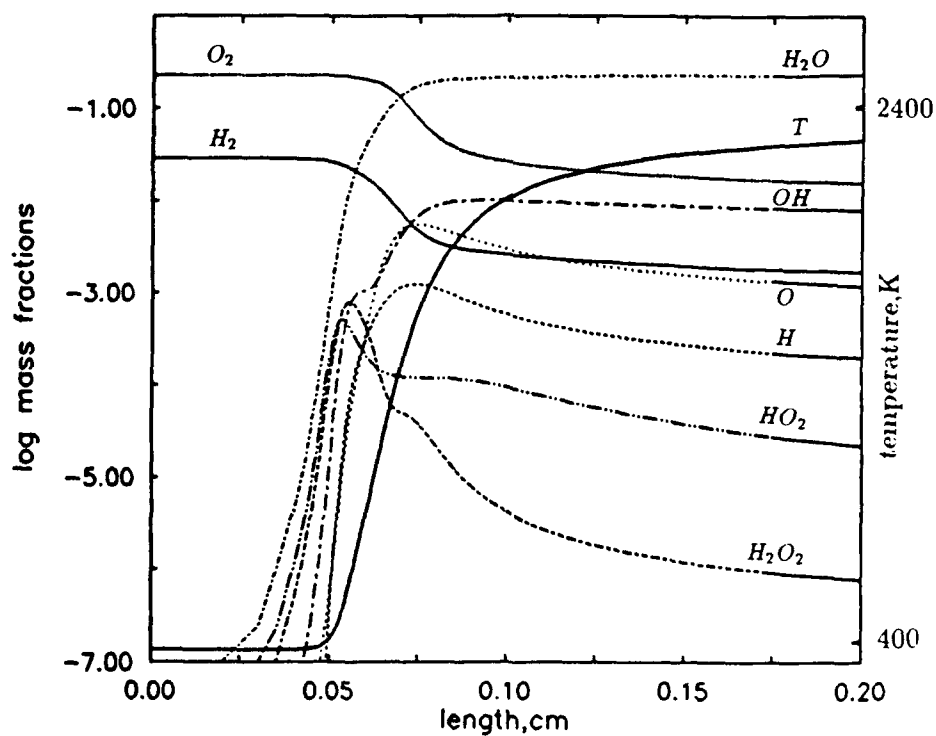


Fig. 10. Mass fractions and temperature for the stoichiometric flame.

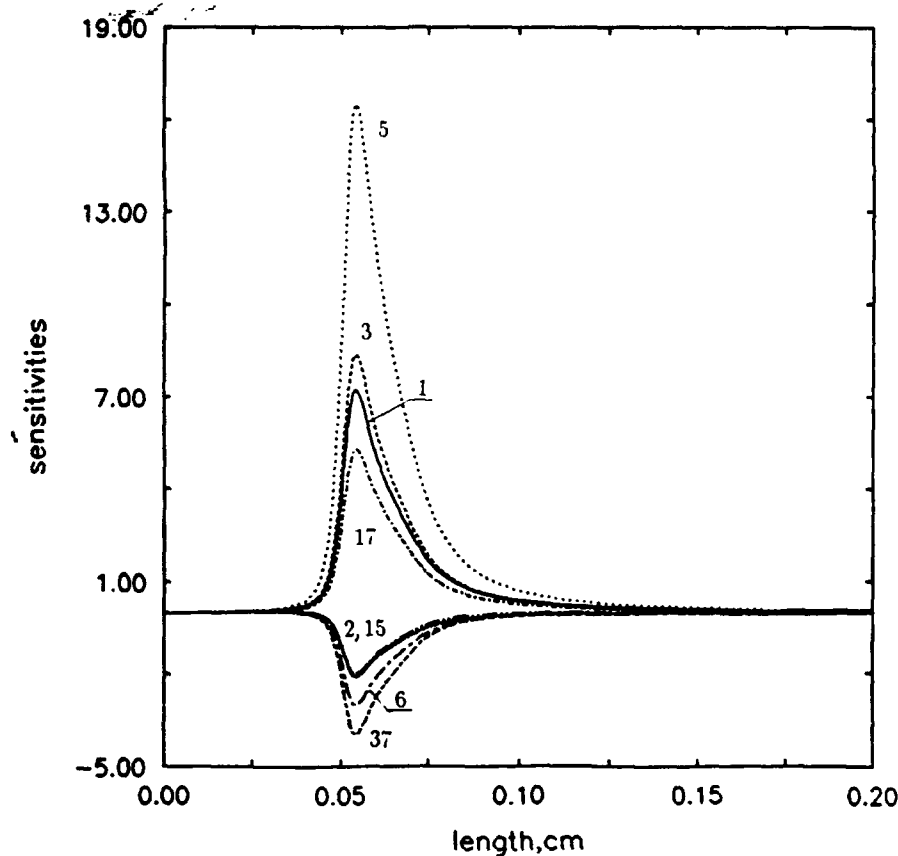


Fig. 11. Normalized sensitivity functions of the temperature for the most important reactions in the stoichiometric flame.

$T = 400$ K at $x = 0.05$ cm. The adiabatic calculation yields the flame speed $u = 236.5$ cm s⁻¹, which is in the range of experimental data (see Refs. 21 and 42 for reviews), though slightly higher than the one computed in Ref. 35. Comparing Figs. 4 and 10 we might assume that molecular and thermal diffusion merely "smooth" the abrupt changes in temperature and mass fraction profiles. Principal component analysis based on the sensitivities of all species and the temperature shows a similar "smoothing" effect on the relative importance of elementary reactions. Although the sensitivity functions for steps 1 and 17 are much smaller than in the adiabatic, diffusion-free calculations, a large number of reactions is at least slightly influential for some of the species. Figure 11 shows the temperature sensitivity functions for the most important reactions 5, 3, 1, 17, 37, 6, 2, and 15. The maximum sensitivities are, however, almost as large for steps 21, 19, 4, 7, and 16, not shown in Fig. 11.

According to Fig. 11, the temperature is sensi-

tive to the parameters only in a neighborhood of the flame sheet. This interval is larger than in Fig. 6 for the adiabatic case, and we see that the temperature sensitivity functions are similar. The most remarkable property found in the flame calculation is that there exists the same similarity between the sensitivity functions of all species, and not only of the temperature. For example, Fig. 12 shows the sensitivity functions of the H_2 mass fraction with respect to the Arrhenius parameter A_j of the most important reactions. Although the form of the sensitivity functions is different for each of the species, it is almost independent of the parameter being perturbed, and we have similarly "regular" plots for each species. Thus, by appropriately scaling the sensitivity functions of a species, they will approximately fit on a single curve. For historical reasons the similarity among various variables of a dynamic system (in this case, the sensitivity equations) is frequently called self-similarity [27]. The presence of self-similarity has been demonstrated

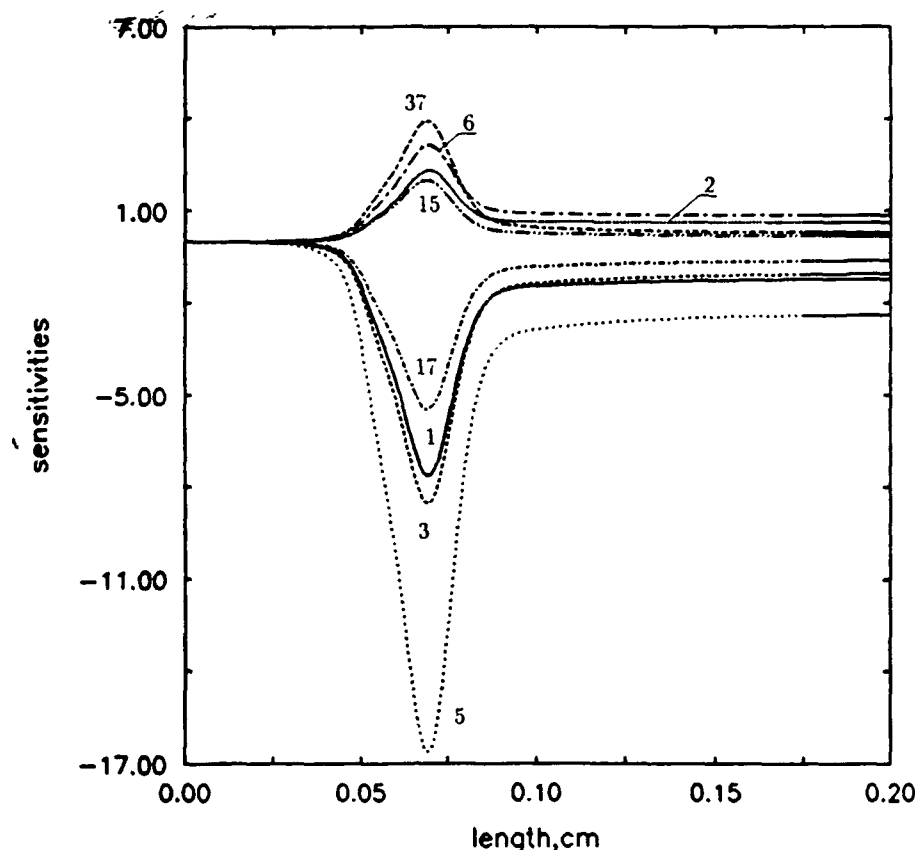


Fig. 12. Normalized sensitivity functions of the H_2 mass fraction for the most important reactions in the stoichiometric flame.

in several combustion systems, including numerical testing of the relationships [27].

We can use the decomposition introduced in the previous section to show that the similarity of mass fraction sensitivity functions follows from the similarity of temperature sensitivity functions, but our derivation should be slightly generalized.

For notational simplicity we write Eqs. 3 and 4 abstractly as

$$L_k(Y_1, \dots, Y_K, T, \mathbf{p}) = 0, \quad k = 1, \dots, K \quad (20)$$

and

$$L_T(Y_1, \dots, Y_K, T, \mathbf{p}) = 0, \quad (21)$$

where L_k , $k = 1, \dots, K$, and L_T denote the second-order differential operators in Eqs. 3 and 4, respectively. The corresponding boundary conditions are given by Eqs. 6 and 7. As in the

previous section we write the K equations in Eq. 20 as a vector equation

$$\mathbf{L}(\mathbf{Y}, T, \mathbf{p}) = 0, \quad (22)$$

where $\mathbf{L} = (L_1, \dots, L_K)^T$ and $\mathbf{Y} = (Y_1, \dots, Y_K)^T$. The sensitivity functions of interest are $\partial Y_i(x, \mathbf{p}) / \partial p_j$, $i = 1, \dots, K$, and $\partial T(x, \mathbf{p}) / \partial p_j$ that form the sensitivity matrix $\partial \mathbf{Y} / \partial \mathbf{p}$ and vector $\partial T / \partial \mathbf{p}$, respectively. By differentiation of Eq. 22 these coefficients satisfy the sensitivity equation

$$\left(\frac{\partial \mathbf{L}}{\partial \mathbf{Y}} \right)_x \frac{\partial \mathbf{Y}}{\partial p_j} + \left(\frac{\partial \mathbf{L}}{\partial T} \right)_x \frac{\partial T}{\partial p_j} + \left(\frac{\partial \mathbf{L}}{\partial \mathbf{p}_j} \right)_x = 0, \quad (23)$$

where $(\partial \mathbf{L} / \partial \mathbf{Y})_x$, $(\partial \mathbf{L} / \partial T)_x$, and $(\partial \mathbf{L} / \partial \mathbf{p}_j)_x$ are differential operators. Similarly to the previous section, we express the sensitivity coefficients

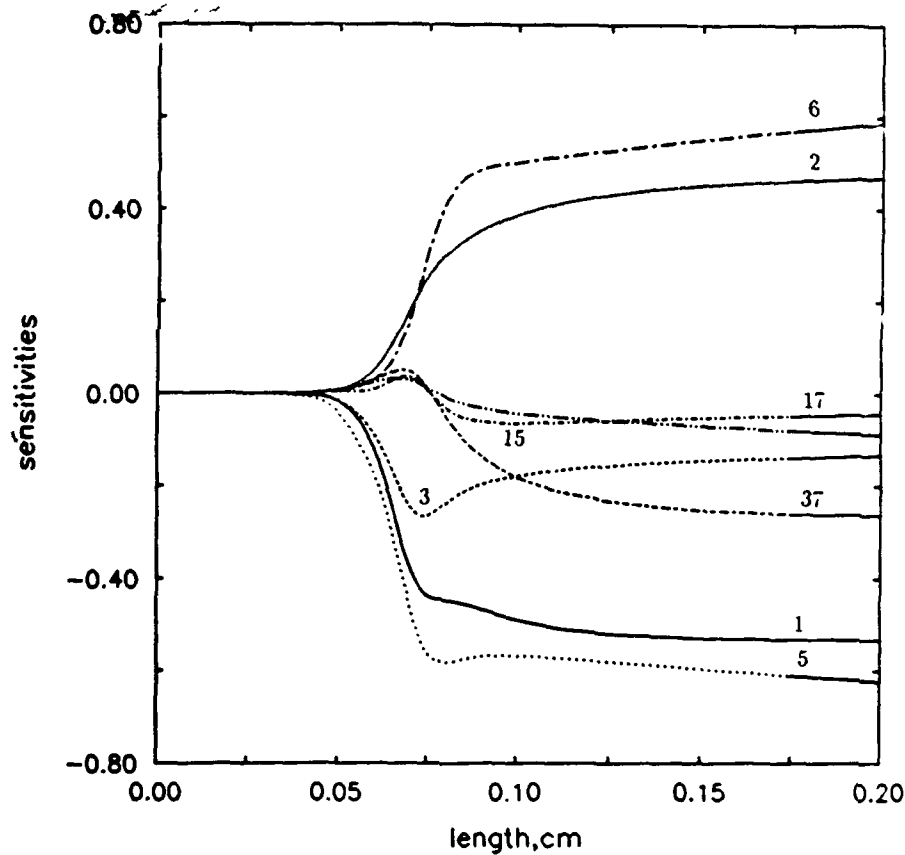


Fig. 13. Constrained temperature sensitivity functions of the H_2 mass fraction for the reactions considered in Figs. 11 and 12.

$\partial Y/\partial p_j$ in terms of the $K \times K$ Green's function matrix $G_1(x, x')$ of the system (Eq. 22), where $G_1(x, x')$ is defined to satisfy the equation

$$\left(\frac{\partial L}{\partial Y} \right)_x G_1(x, x') = -I \delta(x - x'). \quad (24)$$

For convenience we assume that the Green's function satisfies the same boundary conditions as the sensitivity coefficients $\partial Y/\partial p_j$ in Eq. 23. Because there are two inhomogeneous terms in Eq. 23, we have the decomposition

$$\begin{aligned} \frac{\partial Y}{\partial p_j}(x) &= \int_0^L G_1(x, x') \left(\frac{\partial L}{\partial T} \right)_{x'} \\ &\quad \times \frac{\partial T}{\partial p_j}(x') dx' + \int_0^L G_1(x, x') \\ &\quad \times \left(\frac{\partial L}{\partial p_j} \right)_{x'} dx'. \end{aligned} \quad (25)$$

The second term in Eq. 25 stands for the constrained temperature sensitivity functions that are the solutions of Eq. 23 with $\partial T/\partial p_j = 0$ and the adiabatic flame temperature profile as a parameter-independent external variable. Figure 13 shows the constrained temperature sensitivity functions of the H_2 mass fraction for the same reactions whose flame sensitivity functions are shown in Fig. 12. According to these plots, there exists a neighborhood $[x_1, x_2]$, $0 < x_1 < x_2 < L$, of the flame sheet such that the first term in Eq. 25 is dominant on $[x_1, x_2]$, whereas the sensitivities are small outside this interval. Therefore, for any $x \in [x_1, x_2]$ we have

$$\frac{\partial Y}{\partial p_j} \approx \int_0^L G_1(x, x') \left(\frac{\partial L}{\partial T} \right)_{x'} \frac{\partial T}{\partial p_j}(x') dx'. \quad (26)$$

We refer to Eq. 26 as the strong coupling approximation of the sensitivity functions, based on the

property that the parameter perturbations influence the mass fraction profiles almost exclusively through the change induced in the temperature profile.

The validity of the strong coupling approximation (Eq. 26) enables us to understand why the self-similarity of the temperature sensitivity functions shown in Fig. 12 implies self-similarity of the mass fraction sensitivity functions. Consider two parameters p_i and p_j . By the self-similarity of the corresponding temperature sensitivity functions there exists a constant c such that $\partial T(x, \mathbf{p})/\partial p_j \approx c \partial T(x, \mathbf{p})/\partial p_i$ for all x . Then the linearity of 26 implies $\partial Y(x, \mathbf{p})/\partial p_j \approx c \partial Y(x, \mathbf{p})/\partial p_i$ for all $x \in [x_1, x_2]$, and thus the same similarity holds for the sensitivity functions of all mass fractions. Although Eq. 26 usually is a good approximation, there obviously exist some deviations from perfect similarity due to the neglected second term in Eq. 25. For example, the temperature sensitivity functions for reactions 2 and 15 are almost indistinguishable (see Fig. 11), whereas the sensitivities of the H_2 mass fraction with respect to the same parameters markedly differ, as shown in Fig. 12. The deviation clearly stems from the fact that the constrained temperature sensitivity function of Y_{H_2} for step 2, though relatively small, is much larger than the one for step 15 (see Fig. 13). All deviations from the perfect similarity in Fig. 12 can be similarly explained in terms of the constrained temperature sensitivity functions shown in Fig. 13. The deviations are also clearly shown by principal component analysis. The validity of the strong coupling approximation (Eq. 26) has a number of practical consequences. First, modeling a combustion process with the measured temperature profile as input data should be relatively reliable, since uncertainties in rate coefficients slightly influence the computed mass fractions. Because this conclusion is based on local sensitivity analysis, it cannot be extended to large parameter variations. Conversely, it is important that the sensitivity functions computed for a prescribed temperature profile (see, e.g. Refs. 43 and 44) are not very informative, since eliminating the large first term in Eq. 25 will produce results that do not reflect the true importance of elementary reactions in the

flame. Second, almost all information on kinetic parameters derivable from the observation of a flame is contained in the temperature data or one concentration profile, and relatively little more can be learned from the observations of several concentration variables. This also implies that sensitivity results in a flame can be summarized as r numbers, where r is the number of parameters, for example the values $c_j = (\partial T/\partial p_j)_{\max}$, $j = 1, \dots, r$. Indeed, the same coefficients of proportionality apply to the sensitivity functions of each variable in the system, and hence the numbers c_1, \dots, c_r represent the relative importance of elementary reactions. We emphasize that by neglecting the second term in Eq. 25 this is an approximation, and the sensitivity functions actually contain somewhat more information than can be extracted, for example, by principal component analysis.

Introducing the strong coupling approximation (Eq. 26) we have shown that self-similarity of the temperature sensitivity functions implies self-similarity of sensitivity functions of all the other variables. This is sufficient for the purposes of the present article, but does not explain why the temperature sensitivity functions themselves are self-similar. Because self-similarity has been observed in a number of steady flame calculations [27, 35, 45-46], establishing the mildest assumptions additional to the strong coupling approximation (Eq. 26) deserves further investigation.

Self-Similarity and Mechanism Reduction

In the remainder of the article we exploit self-similarity, in this section for mechanism reduction. Our aim is to find the simplest mechanism that is able to reproduce the "observables," i.e., the flame speed, the temperature, and mass fractions for H_2 , O_2 , and H_2O , within reasonable errors. As discussed in section 4, a potential method of finding such minimal mechanism is restricting considerations to the sensitivity functions of the "observable" variables. The approach is, however, of no general validity, and we failed when trying to further reduce the 23-step mechanism found in the diffusion-free cases.

Due to self-similarity the situation is different

TABLE 3

Principal Components for Stoichiometric Flame, Temperature, and Mole Fractions for the "Observable" Species, All Rate, and Transport Coefficients as Parameters

No.	Eigenvalue ^a	Parameters in the Principal Component ^b
1	2.21(+6)	1[0.21], 3[0.24], 5[0.48], 39[0.56], 40[-0.32], 42[0.39]
2	1.98(+2)	1[-0.33], 3[-0.25], 5[-0.52], 39[0.45], 40[-0.28], 42[0.36]
3	1.38(+2)	46[0.99]
4	1.54(+1)	1[0.35], 2[-0.45], 3[-0.21], 6[-0.50], 37[0.38], 40[-0.25]
5	7.70(+0)	1[-0.21], 5[0.22], 6[-0.31], 17[-0.40], 40[0.67], 42[0.38]
6	3.03(+0)	1[-0.29], 5[0.22], 6[-0.31], 17[-0.40], 40[-0.49], 42[0.37]
7	8.32(-1)	17[0.81], 42[0.32]
8	4.71(-1)	1[0.41], 4[-0.21], 5[-0.21], 6[0.25], 7[-0.21], 8[0.21], 16[0.20], 39[-0.26], 42[0.50]
9	1.24(-1)	2[0.43], 3[0.26], 4[-0.23], 6[-0.40], 7[-0.21], 8[0.29], 21[-0.43]
10	9.57(-2)	1[0.41], 4[0.35], 6[0.25], 15[0.37], 16[-0.39], 37[-0.35], 42[0.21]
11	4.27(-2)	2[0.38], 4[0.23], 5[0.29], 7[-0.27], 8[0.29], 37[0.57], 38[-0.29]

^a Numbers in parentheses denote powers of ten.

^b Numbers in brackets denote the coefficients of the parameters in the corresponding principal component.

in the steady premixed flame. Because the sensitivity functions are similar, we can restrict consideration to a single "observable" when ranking the reactions according to their importance. The ranking will be then valid for all variables up to the accuracy of the strong coupling approximation. This will enable us to reduce the number of reactions. For example, restricting consideration to the temperature in principal component analysis yields the mechanism consisting of Steps 1, 3, 5, 17, 2, 37, 6, 21, and 38, in order of decreasing importance. The strong coupling approximation, i.e., neglecting the second term in Eq. 25 gives, however, up to 10% errors that propagates into the mass fraction profiles computed from the reduced mechanism. Therefore, in addition to the temperature sensitivity function, it is advisable to include also the mass fraction sensitivity functions of H_2 , O_2 , and H_2O in principal component analysis. The parameters considered in this calculation are the Arrhenius parameters A_1, \dots, A_{38} , the thermal conductivity coefficient λ denoted by p_{39} , and the diffusion coefficients D_{H_2} , D_{O_2} , D_H , D_O , D_{OH} , D_{HO_2} , $D_{H_2O_2}$, and D_{N_2} , denoted by p_{40} - p_{48} . Since we have $q = 87$ mesh points in the flame calculation, and consider $m = 4$ observables in the principal component analysis, the cutoff value of the eigenvalues is

$\lambda = mq \times 10^{-4} \approx 3.5 \times 10^{-2}$. The eigenvalues exceeding this threshold and the corresponding principal components are listed in Table 3. There are only 15 rate constants present in these principal components. These reactions will be of importance for further analysis and hence are listed in Table 4.

According to Table 3, in addition to the rate coefficient of the selected 15 reactions, further important parameters are p_{39} , p_{40} , p_{41} , and p_{46} , i.e., the thermal conductivity λ and the diffusion coefficients D_{H_2} , D_H , and D_{H_2O} . The

TABLE 4
Reduced Mechanism of H_2 Oxidation

No.	Reaction
1-2	$H + O_2 \leftrightarrow O + OH$
3-4	$O + H_2 \leftrightarrow H + OH$
5-6	$H_2 + OH \leftrightarrow H_2O + H$
7-8	$O + H_2O \leftrightarrow OH + OH$
15-16	$H + OH + M \leftrightarrow H_2O + M$
17	$H + O_2 + M \leftrightarrow HO_2 + M$
19	$H + HO_2 \rightarrow H_2 + O_2$
21	$H + HO_2 \rightarrow OH + OH$
37-38	$OH + O + M \leftrightarrow HO_2 + M$

TABLE 5

Mechanism Reduction for H₂-Air Flames: Deviations of the Temperature and Mole Fraction Profiles

x (cm)	"Observable"	Stoichiometric		Lean ^a		Rich ^b	
		Complete Mechanism	15 steps, % deviations	Complete Mechanism	15 steps, % deviations	Complete Mechanism	15 steps, % deviations
0.0544	T (K)	615.3	0.11	566.5	0.93	599.2	-0.01
	X_{H_2}	2.312(-1)	1.51	1.747(-1)	3.09	4.623(-1)	0.15
	X_{O_2}	1.501(-1)	-1.60	1.646(-1)	-1.88	1.030(-1)	-1.74
	X_{H_2O}	2.498(-2)	6.61	2.448(-2)	4.82	1.879(-2)	12.16
0.060	T (K)	953.9	-1.68	856.0	-0.65	892.4	-1.19
	X_{H_2}	1.702(-1)	3.81	1.293(-1)	4.87	4.139(-1)	0.82
	X_{O_2}	1.397(-1)	-1.78	1.575(-1)	-2.41	9.251(-2)	-2.11
	X_{H_2O}	7.687(-2)	-3.07	6.667(-2)	-1.53	5.706(-2)	-0.49
0.070	T (K)	1503.0	-2.79	1352.0	-1.85	1307.1	-3.21
	X_{H_2}	6.745(-2)	7.97	5.302(-2)	8.84	3.148(-1)	0.86
	X_{O_2}	7.306(-2)	3.46	1.126(-1)	-0.62	4.114(-2)	-5.10
	X_{H_2O}	2.070(-1)	-3.52	1.673(-1)	-2.57	1.655(-1)	-4.35
0.0	u (cm s ⁻¹)	236.5	7.06	147.9	11.62	356.3	7.01

^a Lean mixture boundary conditions: $X_{H_2} = 0.2495$, $X_{O_2} = 0.1578$, $X_{N_2} = 0.5927$, where X denotes the mole fractions.^b Rich mixture boundary conditions: $X_{H_2} = 0.5000$, $X_{O_2} = 0.1051$, $X_{N_2} = 0.3949$, where X denotes the mole fractions.

influence of the diffusion of H and H₂ on the combustion rate is well known [2]. The importance of D_{H_2O} , which is not coupled with any of the kinetic parameters according to the principal component, ψ_3 , is somewhat surprising, and likely stems from the large efficiency factor of H₂O (see [M] in Table 1).

The ability of reducing the mechanism is based on the approximation (Eq. 26), which is valid only on some interval $[x_1, x_2]$ containing the flame sheet positioned at $x = 0.0544$ cm for the stoichiometric mixture. The first part of Table 5 shows the values of the "observables" at some points of this interval computed with the full 38-step mechanism and the percent deviations when reducing the mechanism to the 15 steps listed in Table 4. We note that in the same region the deviations are small also for the radicals, H[•], O[•], OH[•], and HO₂[•], whose sensitivities have not been considered in the principal component analysis. As we conjectured in section 4, the system is so strongly coupled that a reduced mechanism

is able to provide good approximations for the molecular species only if the radicals are predicted well. This is the reason why no reduction of the mechanism was possible in the diffusion-free calculations, without the property of self-similarity.

There is no reaction producing H₂O₂ in the reduced mechanism, and hence this species can be omitted. Step 20, the only initiation reaction, is also omitted, since the radicals are mostly supplied by diffusion from the post-flame region. This emphasizes that the reduced mechanism applies only to modeling steady premixed flames, i.e., the same system in which the sensitivities used for reduction have been computed. As shown in the further columns of Table 5, the reduced mechanism gives good prediction for the "observables" also in lean and rich H₂-air flames. The deviations are larger for the flame speed, which were not considered in principal component analysis. We discuss this latter problem in the next section.

TABLE 6

Principal Components for Stoichiometric Flame, Temperature and Mole Fractions for the "Observable" Species, 15 Rate Coefficients of the Reduced Mechanism as Parameters

No.	Eigenvalue ^a	Parameters in the Principal Component ^b
1	9.35(+5)	1[0.32], 3[0.38], 5[0.74], 17[0.24]
2	1.91(+1)	1[-0.44], 2[0.48], 6[0.52], 15[-0.20], 37[-0.38]
3	4.70(+0)	1[0.44], 5[-0.37], 6[0.35], 17[0.69]
4	8.05(-1)	1[0.47], 3[0.27], 6[0.30], 17[-0.66]
5	1.62(-1)	2[0.49], 3[0.27], 4[-0.21], 6[-0.41], 7[0.23], 8[0.30], 19[0.21], 21[-0.46]
6	1.45(-1)	1[0.26], 4[0.37], 7[0.24], 15[0.39], 16[-0.42], 37[-0.50]
7	4.91(-2)	2[-0.32], 4[-0.31], 5[-0.22], 6[-0.24], 15[-0.21], 21[-0.24]
8	3.42(-2)	3[0.45], 7[0.43], 8[-0.37], 16[0.47], 19[0.29], 38[-0.30]
9	1.26(-2)	1[0.23], 3[-0.35], 6[-0.34], 7[-0.22], 8[0.23], 15[-0.22], 16[0.42], 21[0.22], 37[-0.45], 38[-0.30]
10	8.88(-3)	1[0.49], 5[-0.44], 6[-0.30], 16[-0.27], 19[-0.31], 21[0.48]
11	3.92(-3)	1[0.34], 2[0.55], 3[-0.21], 7[0.25], 8[-0.29], 15[-0.28], 19[-0.29], 37[0.21], 38[0.36]
12	2.18(-3)	4[-0.55], 15[0.66], 16[0.28], 21[0.25]
13	7.80(-4)	4[-0.59], 15[-0.32], 16[-0.48], 19[0.38], 21[0.24], 38[-0.26]
14	3.55(-4)	19[0.62], 21[0.44], 38[0.60]
15	6.95(-5)	7[0.71], 8[0.70]

^a Numbers in parentheses denote powers of ten.

^b Numbers in brackets denote the coefficients of the parameters in the corresponding principal component.

Self-Similarity and Kinetic Model Simplification

In kinetic model simplification we introduce further assumptions such as the QSSA to find the simplest possible models that give tolerable errors in flame calculations. For H_2 oxidation there exist a number of very simple empirical models (see, e.g., Ref. 47) that perform relatively well, at least for limited regions of the composition space. It is also known that the QSSA applies to the radicals except H [21]. In this section we try to understand why the simple models work, considering the stoichiometric H_2 -air flame as an example. Our starting point is the 15-step reduced mechanism shown in Table 4. As discussed in section 3, mechanistic interpretation of principal components corresponding to small eigenvalues may help to identify applicable simplifying assumptions [25, 26]. Considering the "observables" T , Y_{H_2} , Y_{O_2} , and Y_{H_2O} , and restricting consideration to the preexponential factors A_j of the 15 reactions, principal components are listed in Table 6. The cut-point for small eigenvalues is $\lambda_{min} \approx 0.035$, and there are seven eigenvalues below this threshold. According to the principal

component ψ_{15} , the "observables" depend only on the ratio k_7/k_8 , and hence the partial equilibrium assumption is expected to apply to this pair of reactions. We emphasize that, based on local sensitivity analysis, any such conclusion should be verified by calculation. The simplest way of testing the validity of the assumption is to multiply A_7 and A_8 by the same large factor. The "observables" are expected to be almost invariant under such move in the parameter space. According to column A of Table 7, the parameters $A_7 = 100 A_7^\circ$ and $A_8 = 100 A_8^\circ$, where A_7° and A_8° denote the original (nominal) values, give rise to relatively small deviations. However, the flame speed, not considered in principal component analysis, is significantly decreased. The second smallest eigenvalue is λ_{14} , and the corresponding principal component ψ_{14} includes A_{19} , A_{21} , and A_{38} , i.e., the rate constants of reactions of the reduced mechanism that consume HO_2 . The only explanation is that steps 17 and 37 are rate determining, and the QSSA applies to HO_2 . The simplest way to check this assumption is increasing the values of A_{19} , A_{21} , and A_{38} by moving along the eigenvector u_{14} . For exam-

TABLE 7

Model Simplification for the Stoichiometric H₂-Air Flame: Deviations of the Temperature and Mole Fraction Profiles

x (cm)	Variable	Complete Mechanism	Reduced Mechanism or Modified Model (% Deviations)						
			A	B	C	D	E	F	G
0.0544	T (K)	615.3K	-2.16	-3.33	-2.59	2.12	3.12	2.65	2.71
	X _{H₂}	2.312(-1)	-2.46	-2.12	-12.50	-2.81	-2.21	1.16	1.12
	X _{O₂}	1.501(-1)	1.93	2.59	5.66	2.46	4.86	4.26	4.26
	X _{H₂O}	2.49(-2)	-6.84	-11.49	24.82	-9.60	-35.07	-48.67	-49.11
0.0600	T (K)	953.8	-0.49	-1.67	6.41	8.08	10.07	8.29	8.39
	X _{H₂}	1.704(-1)	-3.58	-2.64	-15.32	-14.50	-17.60	-9.38	-9.86
	X _{O₂}	1.397(-1)	3.43	5.22	7.02	3.65	9.95	11.31	11.45
	X _{H₂O}	7.687(-2)	-0.73	-4.84	19.47	17.65	14.25	-6.43	-6.81
0.0700	T (K)	1503.0	1.26	0.86	15.10	9.05	8.84	4.39	5.12
	X _{H₂}	6.745(-2)	-8.33	-6.44	-12.34	-43.32	-54.13	-67.36	-75.34
	X _{O₂}	7.306(-2)	-0.33	4.04	16.10	-34.25	-37.44	-27.99	-28.66
	X _{H₂O}	2.070(-1)	2.46	1.26	6.71	17.98	22.80	21.78	23.8
0.0	u (cm s ⁻¹)	236.5	-12.90	-16.65	-32.30	1.39	-2.33	-1.78	-1.18

^a A: $A_7 = 100 A_7^0$, $A_8 = 100 A_8^0$.

B: $A_7 = 100 A_7^0$, $A_8 = 100 A_8^0$, $A_{19} = 10 A_{19}^0$, $A_{21} = 5.12 A_{21}^0$, $A_{38} = 9.28 A_{38}^0$.

C: $A_{17} = 5.27 A_{17}^0$, $A_{37} = 10 A_{37}^0$, $A_{38} = 10 A_{38}^0$.

D: $A_1 = 1.59 A_1^0$, $A_2 = 2.21 A_2^0$, $A_3 = 0.36 A_3^0$, $A_4 = 0.06 A_4^0$, $A_5 = 1.90 A_5^0$, $A_6 = 1.58 A_6^0$, $A_7 = 805.67 A_7^0$, $A_8 = 328.53 A_8^0$, $A_{15} = 1.01 A_{15}^0$, $A_{16} = 0.63 A_{16}^0$, $A_{19} = 83.11 A_{19}^0$, $A_{21} = 16.75 A_{21}^0$, $A_{37} = 1.39 A_{37}^0$, $A_{38} = 19.28 A_{38}^0$.

E: $A_1 = 0.73 A_1^0$, $A_2 = 1.86 A_2^0$, $A_3 = 0.03 A_3^0$, $A_4 = 0.12 A_4^0$, $A_5 = 41.40 A_5^0$, $A_6 = 35.28 A_6^0$, $A_7 = 5469.70 A_7^0$, $A_8 = 667.96 A_8^0$, $A_{15} = 1.39 A_{15}^0$, $A_{16} = 1.14 A_{16}^0$, $A_{19} = 5965.57 A_{19}^0$, $A_{21} = 1.01 A_{21}^0$, $A_{37} = 4.92 A_{37}^0$, $A_{38} = 363.58 A_{38}^0$.

F: $A_1 = 0.51 A_1^0$, $A_2 = 0.59 A_2^0$, $A_3 = 0.00 A_3^0$, $A_4 = 0.00 A_4^0$, $A_5 = 371.37 A_5^0$, $A_6 = 97.71 A_6^0$, $A_7 = 1096.70 A_7^0$, $A_8 = 795.14 A_8^0$, $A_{15} = 0.00 A_{15}^0$, $A_{16} = 0.00 A_{16}^0$, $A_{19} = 3701.37 A_{19}^0$, $A_{21} = 0.00 A_{21}^0$, $A_{37} = 1.68 A_{37}^0$, $A_{38} = 0.00 A_{38}^0$.

G: $A_1 = 0.49 A_1^0$, $A_2 = 1.08 A_2^0$, $A_3 = 0.00 A_3^0$, $A_4 = 0.00 A_4^0$, $A_5 = 5809.81 A_5^0$, $A_6 = 1132.13 A_6^0$, $A_7 = 5698.80 A_7^0$, $A_8 = 1038.25 A_8^0$, $A_{15} = 0.00 A_{15}^0$, $A_{16} = 0.00 A_{16}^0$, $A_{19} = 17493.30 A_{19}^0$, $A_{21} = 0.00 A_{21}^0$, $A_{37} = 4.54 A_{37}^0$, $A_{38} = 0.00 A_{38}^0$.

ple, the values $A_{19} = 10 A_{19}^0$, $A_{21} = 5.12 A_{21}^0$, and $A_{38} = 9.28 A_{38}^0$, in addition to the already increased values of A_7 and A_8 , result in the deviations shown in column B of Table 7. Except the value of X_{H_2O} at the flame sheet, the "observables" are well reproduced, but the flame speed is further decreased.

We would like to further simplify the model and to avoid the deviations in the predictions of the flame speed. The eigenvectors corresponding to further small eigenvalues are, however, too complex for mechanistic interpretation. Before

trying to formulate a more-or-less systematic procedure we emphasize that any simplifying kinetic assumption can be regarded as a move in the parameter space. For example, reactions 7 and 8 are in partial equilibrium if and only if we can increase k_7 and k_8 arbitrarily, while keeping their ratio fixed at the equilibrium constant. Similarly, HO_2 is in quasi-steady state if and only if the rates of reactions consuming it can be arbitrarily increased, possibly keeping their ratios fixed. Therefore, in order to exploit the sensitivity results for model simplification, we look for

such "invariant" directions in the parameter space, trying to preserve the value of the flame speed at the same time.

Let $Y_i(x, \mathbf{p})$ denote the i th "observable." From the Taylor series expansion the deviation $\Delta Y_i(x, \mathbf{p}) = Y_i(x, \mathbf{p}) - Y_i(x, \mathbf{p}^\circ)$ is given by

$$\Delta Y_i(x, \mathbf{p}) = \sum_{j=1}^r \frac{\partial Y_i(x, \mathbf{p}^\circ)}{\partial p_j} (p_j - p_j^\circ) + \sigma(\|\Delta \mathbf{p}\|^2), \quad (27)$$

where $\sigma(\|\Delta \mathbf{p}\|^2)$ denotes the higher-order terms. Due to self-similarity, there exist constants $\beta_j = c_j/c_1$, $j = 2, \dots, r$, such that $\partial Y_i(x, \mathbf{p}^\circ)/\partial p_j \approx \beta_j \partial Y_i(x, \mathbf{p}^\circ)/\partial p_1$. Therefore, we can select vectors $\Delta \mathbf{p} = \mathbf{p} - \mathbf{p}^\circ$ in infinitely many different ways to make the sum in Eq. 27 vanishingly small. Calculations show, however, that by the presence of higher-order terms $\sigma(\|\Delta \mathbf{p}\|^2)$ this consideration does not enable us to find invariant directions in the parameter space. For example, selecting $A_{37} = 10 A_{37}^\circ$, $A_{38} = 10 A_{38}^\circ$, and $A_{17} = 5.272 A_{17}^\circ$, by self-similarity we have $(\partial Y_i/\partial A_{17})(A_{17} - A_{17}^\circ) + (\partial Y_i/\partial A_{38})(A_{38} - A_{38}^\circ) + (\partial Y_i/\partial A_{37})(A_{37} - A_{37}^\circ) \approx 0$ for all i and x , i.e., the effect of changing A_{37} and A_{38} can be compensated by multiplying also A_{17} by a suitable factor. These relatively small perturbations of the parameters give, however, large deviations not only in the flame speed, but also in the values of the "observables," as shown in Column C of Table 7. These deviations are caused clearly by the higher-order terms in Eq. 27.

We emphasize that principal component analysis approximates also the second-order sensitivity functions while requiring only the first-order ones (see Ref. 25) and hence may perform better. According to Table 6 we have seven small eigenvalues, and any parameter perturbation $\Delta \mathbf{p}$ confined to the seven-dimensional subspace spanned by the corresponding eigenvectors is expected to lead to small changes in the "observables," at least when $\|\Delta \mathbf{p}\|$ is not too large. In addition, we want to keep the flame speed u unchanged, and hence restrict consideration to parameter per-

turbations satisfying the equation

$$\sum_{j=1}^r (\partial u/\partial p_j) \Delta p_j = 0. \quad (28)$$

Taking into account this constraint, we still have a six-dimensional subspace of the parameter space to explore, and hence there exist infinitely many "invariant" directions. To find such vectors we consider parameter perturbations and find their projections onto the invariant subspace by least squares method, subject to the flame speed constraint.

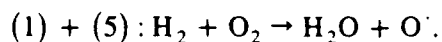
The first result following from this approach is that $\Delta \mathbf{p}_{17}$ is orthogonal to the invariant subspace, and hence it is not possible to change the value of A_{17} without introducing large deviations in the observations. Therefore, in spite of its relatively small sensitivity coefficient, step 17 plays an important role also in flame modeling. This immediately explains the large deviations in column C of Table 7.

We look for parameter perturbations that can be given some mechanistic interpretations in terms of simplifying assumptions. First, we try to increase A_7 and A_8 in order to confirm the partial equilibrium assumption, as well as to increase A_{19} , A_{21} , and A_{38} , thereby moving HO_2 toward its steady-state values. The selected parameters in the invariant subspace and the resulting deviations are shown in column D of Table 7. It follows that subject to the constraint on the flame speed we cannot multiply A_7 and A_8 by the same factor, and hence the partial equilibrium assumption does not globally apply. This agrees with the result of Dixon-Lewis [21], who emphasized that such assumptions are valid only in the recombination region. We show, however, that the QSSA on O , OH and HO_2 radicals is a reasonable global assumption. According to column E of Table 7, increasing also the values of A_5 and A_6 , the deviations are almost unchanged, except the one for $X_{\text{H}_2\text{O}}$ at the flame sheet. Columns F and G show the results of further increasing the parameters. Notice that the rates of reaction 3, 4, 15, 16, 21, and 38 are becoming very small at the same time, since their effects are compensated by increasing the values of A_5 ,

A_6 , A_7 , A_8 , and A_{19} . According to columns F and G of Table 7, the deviations are almost unchanged under very large parameter perturbations in the last step.

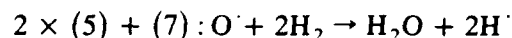
The reason behind these parameter changes will be clear looking at the resulting mechanism of the nine reactions 1, 2, 5, 6, 7, 8, 17, 19, and 37. Because we increased A_5 , A_6 , A_7 , and A_8 by several orders of magnitude, the radicals OH^\cdot and O^\cdot produced in step 1 quickly react in steps 5 and 7, respectively. Therefore, the OH^\cdot and O^\cdot concentrations are small, and the QSSA certainly applies. Similarly, HO_2^\cdot produced in step 17 quickly recombines in step 19, thereby supporting the QSSA also for HO_2^\cdot .

Because the validity of quasi-steady-state assumptions is clear for the model with some of the rates highly increased, and the increase of these rates gives small deviations in the flame speed and the temperature, the same assumptions apply to the original model. We admit that this reasoning is somewhat indirect. In fact, for kinetic models without diffusion the principal component analysis often reveals the simplifying assumptions unambiguously [25, 26]. In the flame problem, however, the reactions are so strongly coupled that we have an entire invariant subspace instead of some well defined and easily interpretable invariant directions. Therefore, we actually had to move in the parameter space to find such interpretable directions. This emphasizes that the perturbations selected are not at all unique, and the model in flame calculations can be simplified in many different ways. For example, steps 3, 4, 15, 16, and 21 are influential at nominal parameters values, and we could drop them only by increasing the rates of some other reactions. With these arbitrary rates, the reactions 1, 2, 5, 6, 7, 8, 17, 19, and 37 do not form a valid mechanism. Nevertheless, this simplification is advantageous for several reasons. First, it is easy to see how the combustion proceeds in the flame. Adding steps 1 and 5 gives the formal equation



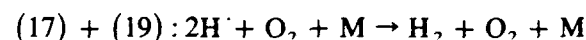
Thus these two steps play the role of the initiation reaction in the presence of H^\cdot radicals, supplied

by diffusion. The rate-determining step is 1. The O^\cdot radical then quickly reacts in the chain-branching reaction 7, producing OH^\cdot radicals. With this additional source of OH^\cdot , step 1 will no more constrain the rate of reaction 5, and the formal reaction

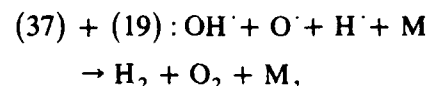


is responsible for the fast increase in the concentration of the radical pool, and for the fast accumulation of the product H_2O . Since $\Delta H_5 = -15$ kcal/mol and $\Delta H_7 = 16.9$ kcal/mol, the sequence $2 \times (5) + (7)$ is exothermic.

The further reactions of the nine-step reduced model form the recombination sequences



and



where $(17) + (19)$ is important only in the low-temperature region.

The simple model also facilitates the derivation of global reaction rates. The quasi-steady-state conditions on O^\cdot , OH^\cdot , and HO_2^\cdot , respectively, are given by

$$\begin{aligned} r_1 - r_2 - r_7 + r_8 - r_{37} &= 0, \\ r_1 - r_2 - r_5 + r_6 + 2r_7 - 2r_8 - r_{37} &= 0, \\ r_{17} + r_{37} - r_{19} &= 0, \end{aligned} \quad (29)$$

where r_i denotes the rate of the i th reaction. Then the production rates of the further species are

$$\begin{aligned} \dot{\omega}_{\text{H}} &= 2R_1 - 2R_2, \\ \dot{\omega}_{\text{H}_2} &= -3R_1 + R_2, \\ \dot{\omega}_{\text{O}_2} &= -R_1, \\ \dot{\omega}_{\text{H}_2\text{O}} &= 2R_1, \end{aligned} \quad (30)$$

where R_1 and R_2 are the global reaction rates defined by

$$R_1 = r_7 - r_8 = \frac{1}{3}(r_5 - r_6) = r_1 - r_2 - r_{37} \quad (31)$$

and

$$R_2 = r_{17} + r_{37} = r_{19}. \quad (32)$$

It follows from the stoichiometry in Eq. 30 that the global reactions are



and



with rate expressions 31 and 32, respectively.

Introduce the notations $u = [\text{OH} \cdot]_{\text{QSS}}$, $v = [\text{O} \cdot]_{\text{QSS}}$, and $z = [\text{HO}_2 \cdot]_{\text{QSS}}$. Rearranging Eqs. 29 gives the quasi-steady-state conditions

$$\begin{aligned} k_7[\text{H}_2\text{O}]v - k_8u^2 + (k_{37}[\text{M}] + k_2)uv \\ = k_1[\text{O}_2][\text{H} \cdot], \end{aligned} \quad (35)$$

$$a = \frac{1}{3} \frac{k_5}{k_8} [\text{H}_2],$$

$$b = \frac{(k_5k_7[\text{H}_2] - k_6(k_{37}[\text{M}] + k_2)[\text{H} \cdot])[\text{H}_2\text{O}]}{3k_8(k_{37}[\text{M}] + k_2)}, \quad (39)$$

$$c = - \frac{k_7[\text{H}_2\text{O}][\text{H} \cdot](3k_1[\text{O}_2] + k_6[\text{H}_2\text{O}])}{3k_8(k_{37}[\text{M}] + k_2)}.$$

Although Eq. 38 can be solved analytically by the Cardano' formula, it does not give a really simple expression for the global reaction rates. Notice that Eq. 38 may have three different real solutions and hence the possibility of flame multiplicity [48] is not excluded in certain concentration regions, but we do not further study this problem here.

Because there are no terms in Eqs. 35 and 36 much smaller than the others, no further simplification of the rate expressions is possible for the entire combustion process. From Eqs. 35 and 36 we have

$$\begin{aligned} 3(k_{37}[\text{M}] + k_2)uv + k_5[\text{H}_2]u \\ = (3k_1[\text{O}_2] + k_6[\text{H}_2\text{O}])[\text{H} \cdot]. \end{aligned} \quad (40)$$

In the preflame region and in the flame sheet we may assume that the recombination is negligible,

$$\begin{aligned} -3k_7[\text{H}_2\text{O}]v + 3k_8u^2 + k_5[\text{H}_2]u \\ = k_6[\text{H}_2\text{O}][\text{H} \cdot], \end{aligned} \quad (36)$$

and

$$k_{19}[\text{H} \cdot]z = k_{17}[\text{H} \cdot][\text{O}_2][\text{M}] + k_{37}uv[\text{M}]. \quad (37)$$

Equations 35 and 36 enable us to find the steady-state radical concentrations u and v , and then z is given by Eq. 37. Unfortunately, in spite of the highly simplified model, Eqs. 35 and 36 give rise to the cubic equation

$$u^3 + au^2 + bu + c = 0, \quad (38)$$

where

and hence $3(k_{37}[\text{M}] + k_2)v \approx k_5[\text{H}_2]$. Then by Eq. 40 the global rate expressions are given by

$$R_1 = k_1[\text{H} \cdot][\text{O}_2], \quad R_2 = k_{17}[\text{H} \cdot][\text{O}_2]. \quad (41)$$

Thus in these regions the most important process is the competition of steps 1 and 17, similarly to the diffusion-free situation. The recombination reactions are, however, important at latter stages of the process.

Although the QSSA on $\text{HO} \cdot$, $\text{O} \cdot$, and $\text{HO}_2 \cdot$ applies also to lean and rich H_2 -air flames, to derive this result from the sensitivity coefficients we had to construct sequences of models, converging to the quasi-steady-state one, that differ from the ones reported in Table 7 for the stoichiometric flame. Because the conditions 35-37 involve the corrupted rate constants of the 9-step model derived for the stoichiometric flame, to

obtain a more generally valid quasi-steady-state model one has to consider the 15-step mechanism in Table 4 as the starting point, thereby obtaining more complex QSSA conditions.

We admit that some results of this section are negative. First, the global rate equations obtained by QSSA on the radicals O^{\cdot} , OH^{\cdot} , and HO_2^{\cdot} do not have a really simple analytic form. Second, because we have too much freedom in simplifying the model, principal component analysis does not directly reveal how to actually perform the simplification and hence it does not offer a practical method. Third, it follows from Table 7 that the simplified model, while predicting the temperature and the flame speed, leads to significant deviations in the mass fraction profiles. We have shown, however, that the combustion mechanism, very complex in a diffusion-free system, is rendered much simpler by the presence of diffusion.

CONCLUSIONS

Although the primary goal of this work is to study the influence of heat release and diffusion on the relative significance of elementary reactions in the mechanism of H_2 oxidation, results help us to understand why highly simplified models can be used in premixed steady flame calculations in spite of an inherently complex reaction mechanism. The complexity of the mechanism of H_2 oxidation has been shown by performing first isothermal, diffusion-free calculations. Sensitivity and principal component analysis reveals that most reactions of our 38-step starting mechanism are influential, and no simplifying kinetic assumptions such as quasi-steady-state on some of radicals apply under these conditions. The influence of thermal effects only has been studied by modeling adiabatic, diffusion-free combustion. Although the feedback through heat release increases the magnitude of sensitivity functions considerably, the conclusions are similar to the isothermal case. Sensitivity functions have also been computed at constrained adiabatic temperature profile, i.e., considering the temperature as an external variable, independent of parameter perturbations. Comparing the two sets of sensitiv-

ity functions it was shown that the indirect effects through thermal feedback are responsible for 70% of the concentration changes brought about by parameter variations. The direct, quasi-isothermal effects are, however, not negligible, and this is a possible explanation of the complexity of the diffusion-free process.

Diffusion has also been considered in the third set of calculations, modeling steady premixed flames. Simultaneous effects of thermal and transport phenomena are shown to change the sensitivity functions dramatically and to lead to their self-similarity. In particular, in the presence of thermal and molecular diffusion the indirect effects of the heat release are responsible for at least 90% of concentration changes brought about by parameter variations. This has been shown by repeating sensitivity calculations with a constrained flame temperature profile. The main consequence is that the concentration of any species is sensitive to the rate constant of a particular reaction if and only if this reaction has a large temperature sensitivity coefficient. This fact enables us to reduce the original mechanism to a set of 15 reactions, thereby introducing less than 5% changes in the concentration and temperature profiles.

By virtue of self-similarity of sensitivity functions, the elementary reactions in the flame model are not kinetically independent, i.e., the effect of changing the rate constant of one reaction can be well compensated by changing the rates of other reactions. Parameter perturbations can be associated with simplifying kinetic assumptions. For example, the ability of increasing the rates of a forward/backward reaction pair while keeping their ratio fixed and thereby introducing only small changes in the solution of the flame model indicates that partial equilibrium of this reaction is a valid assumption. These considerations show that one has much freedom in simplifying the kinetic model in flame calculations. In particular, any parameter perturbation confined to a 6-dimensional subspace of the 15-dimensional parameter space of the reduced mechanism gives relatively small changes in the flame speed and temperature profile. Although the model can be simplified in many different ways, we constructed a

sequence of models for the stoichiometric H_2 -air flame that converge to a 9-step mechanism with quasi-steady-state assumptions on all radicals except H , thereby resulting in a two-step quasi-global model.

The presence of both thermal and molecular diffusion in a steady, one-dimensional flame allow for a much greater degree of simplification than for the case in which one or both of these effects are either absent or can be neglected. In explaining the suitability of simplified models for understanding combustion phenomena, this result is of definitive theoretical interest. Although reduced mechanisms are expected to be practically more valuable for the modeling of multidimensional, nonsteady flames, the mechanism derived here for steady, one-dimensional flames may not be directly transferable to other flame conditions. More generally, the use of any simplified model is restricted to a certain region of the controlling variables in which it can provide an adequate representation of the variables of interest. The methods of sensitivity and principal component analyses are, however, transferable and in conjunction with flame simulations under the appropriate conditions may result in suitable reduced models. It is clear that the relative importance of a particular reaction path significantly depends on these conditions, and the level of possible simplification is influenced also by accuracy requirements.

The authors wish to acknowledge the Office of Naval Research and the Air Force Office of Scientific Research for support of this research.

REFERENCES

1. Benson, S. W., *The Foundations of Chemical Kinetics*, McGraw-Hill, New York, 1960.
2. Westbrook, C. K., and Dryer, F. L., *Prog. Ener. Combust. Sci.* 10:1-57 (1984).
3. Williams, F. A., *Combustion Theory*, Addison-Wesley, Reading, MA, 1965.
4. Buckmaster, J. D., and Ludford, G. S. S., *Theory of Laminar Flames*, Cambridge University Press, Cambridge, 1982.
5. Ludford, G. S. S. (Ed.), *Reacting Flows. Combustion and Chemical Reactors*, North-Holland, Amsterdam, 1986.
6. Ludford, G. S. S. (Ed.), *Lectures in Applied Mathematics*, American Mathematical Society, Providence, 1986, Vol. 24.
7. Peters, N., in *Numerical Simulation of Combustion Phenomena* (R. Glowinski, B. Larrouturou, and R. Temam, Eds.), Lecture Notes in Physics 241, Springer-Verlag, Berlin, 1985, p. 90.
8. Peters, N., and Kee, R. F., *Combust. Flame* 68:17-29 (1987).
9. Fife, P. C., and Nicolenko, B., in *Lectures in Applied Mathematics*, American Mathematical Society, Providence, RI, 1986, Vol. 24, p. 311.
10. Farrow, L. A., and Edelson, D., *Int. J. Chem. Kinet.* 6:787-800 (1974).
11. Farrow, L. A., and Graedel, T. E., *J. Phys. Chem.* 81:2480-2483 (1977).
12. Sundaram, K. M., and Froment, G. F., *Int. J. Chem. Kinet.* 10:1189-1193 (1978).
13. Nicholson, A. J. C., *Can. J. Chem.* 61:1831-1837 (1988).
14. Edelson, D., *Int. J. Chem. Kinet.* 11:687-691 (1979).
15. Rice, O. K., *J. Phys. Chem.* 64:1851-1857 (1960).
16. Benson, S. W., *J. Chem. Phys.* 20:1605-1612 (1952).
17. Bowen, F. R., Acrivos, A., and Oppenheim, A. K., *Chem. Eng. Sci.* 18:177-188 (1963).
18. Volk, L., Richardson, W., Law, K. H., Hall, M., and Lin, S. H., *J. Chem. Educ.* 54:96-97 (1977).
19. Come, G. M., *J. Phys. Chem.* 81:2560-2563 (1977).
20. Klonowski, W., *Biophys. Chem.* 18:73-87 (1983).
21. Dixon-Lewis, G., *Phil Trans. R. Soc. Lond.* 292:45-99 (1979).
22. Dixon-Lewis, G., in *Combustion Chemistry* (W. C. Gardiner, Jr., Ed.), Springer-Verlag, Berlin, 1984.
23. Rabitz, H., in *Reacting Flows. Combustion and Chemical Reactors* (G. S. S. Ludford, Ed.), North Holland, Amsterdam, 1986, p. 67.
24. Rabitz, H., in *Lectures in Applied Mathematics* (G. S. Ludford, Ed.), American Mathematical Society, Providence, RI, Vol. 24, p. 499.
25. Vajda, S., Valko, P., and Turanyi, T., *Int. J. Chem. Kinet.* 17:55-81 (1985).
26. Vajda, S., and Turanyi, T., *J. Phys. Chem.* 90:1664-1669 (1986).
27. Rabitz, H., and Smooke, M. D., *J. Phys. Chem.* 92:1110-1119 (1988).
28. Baulch, D. L., Drysdale, D. D., Horne, D. G., and Lloyd, A. C., *Evaluated Kinetic Rate Data for High Temperature Reactions*, Butterworths, London, 1973.
29. Kee, R. J., Miller, J. A., and Jefferson, T. H., Sandia National Laboratories Report SAND80-8003, 1980.
30. Yetter, R. A., Dryer, F. L., and Rabitz, H., A comprehensive reaction mechanism for carbon monoxide/hydrogen/oxygen kinetics. *Combust. Sci. Technol.* (in press).
31. Dougherty, E. P., and Rabitz, H., *J. Chem. Phys.* 72:6571-6586 (1980).
32. *JANAF Thermochemical Tables*, U.S. National Bureau of Standards Publication NSRDS-NBS37 and sup-

- plements (D. R. Stull and H. Prophet, Eds.), NBS, Washington, DC.
33. Smooke, M. D., *J. Comp. Phys.* 48:72-87 (1982).
 34. Smooke, M. D., Miller, M. D., and Kee, J. F., *Combust. Sci. Technol.* 34:79-89 (1983).
 35. Smooke, M. D., Rabitz, H., Reuven, Y., and Dryer, F. L., Application of sensitivity analysis to premixed hydrogen-air flames, *Combust. Flame* 59:295 (1988).
 36. Gottwald, B. A., and Wanner, G., *Simulation* 37:1969-1975 (1982).
 37. Valko, P., and Vajda, S., *Comput. Chem.* 8:225-271 (1985).
 38. Coffee, T. P., and Heimerl, F. M., *Combust. Flame* 50:323-340 (1983).
 39. Edelson, D., and Allara, D. L., *Int. J. Chem. Kinet.* XII:605-621 (1980).
 40. Hwang, J. T., Dougherty, E. P., Rabitz, S., and Rabitz, H., *J. Chem. Phys.* 69:5180-5191 (1978).
 41. Rabitz, H., *Comput. Chem.* 5:167-181 (1981).
 42. Warnatz, J., *Ber. Bunsenges. Phys. Chem.* 82:643-652 (1978).
 43. Olsson, J. O., Olsson, I. B. M., and Andersson, L. L., *J. Phys. Chem.* 91:4160-4165 (1987).
 44. Olsson, F. O., and Andersson, L. L., *Combust. Flame* 67:99-109 (1987).
 45. Reuven, Y., Smooke, M. D., and Rabitz, H., *J. Comput. Phys.* 64:27-55 (1986).
 46. Mishra, M., Yetter, R., Reuven, Y., Rabitz, H., and Smooke, M. D., Sensitivity analysis of a steady-state premixed laminar CO + H₂ + O₂ flame (in press).
 47. Varma, A. K., Chatwani, A. U., and Bracco, F. V., *Combust. Flame* 64:233-236 (1986).
 48. Clavin, P., Fife, P., and Nicolaenko, B., *SIAM J. Appl. Math.* 47:296-331 (1987).

Received 3 March 1989; revised 28 September 1989

Appendix B

2. Parametric Sensitivity Analysis and Self-Similarity in Thermal Explosion Theory, S. Vajda and H. Rabitz, Chem. Eng. Sci., submitted.

**PARAMETRIC SENSITIVITY AND SELF-SIMILARITY
IN THERMAL EXPLOSION THEORY**

Sandor Vajda

*Department of Biomedical Engineering
Boston University, 44 Cummington Street
Boston, MA 02215*

and

Herschel Rabitz*

*Department of Chemistry
Princeton University
Princeton, NJ 08544*

* To whom correspondence should be addressed

Phone: (609) 258-3917

Submitted to Chem. Eng. Sci., 3/91

Abstract - We study the relations between thermal runaway (also called parametric sensitivity) and self-similarity, an interesting property of the sensitivity functions that has been numerically verified in many explosion and combustion systems. Both concepts are sensitivity-related but independent of the particular parameter being perturbed. This independence is emphasized by proposing a new generalized condition for parametric sensitivity. Criticality is defined as the point in the parameter space where the nominal trajectory exhibits maximum sensitivity to arbitrary, unstructured perturbations applied at the maximum temperature. The condition for criticality reduces to the analysis of the eigenvalues of the Jacobian matrix. In addition to its conceptual generality, the new condition shows that in certain cases there exists no critical Semenov number. The sensitivity functions are shown to satisfy self-similarity relations if and only if the system exhibits critical or supercritical behavior. The onset of self-similarity is explained in terms of two properties of explosion systems, both related to parametric sensitivity. First, the temperature is a dominant variable, and any perturbation in the system affects the conversion mainly through the changes induced in the temperature. This strong coupling of the variables is shown by decomposing the sensitivity functions into direct and indirect terms. Second, the sensitivity equations are pseudohomogeneous in a characteristic time window, in which the system becomes relatively insensitive to parameter perturbations applied within the same interval. The two properties are shown to imply self-similarity of the sensitivity functions. Relations to earlier parametric sensitivity and self-similarity conditions are discussed.

1. INTRODUCTION

This paper is a simultaneous study of two apparently unrelated phenomena. The first is parametric sensitivity or thermal runaway (Morbidelli and Varma; 1988), the second is the self-similarity relation among parameter sensitivity functions, observed in many dynamical systems (Rabitz and Smooke; 1988). We will show that the two concepts are related and the analysis of such relations leads to considerable new insight.

Although both parametric sensitivity and self-similarity are important in a variety of contexts, we restrict consideration to the simple case of a homogeneous system in which an exothermic, irreversible n th order reaction occurs. As shown by Boddington et al. (1983), such a system can be described by the following dimensionless mass and heat balance equations:

$$\frac{dz}{d\tau} = \frac{\psi}{B}(1-z)^n h(\theta) \quad (1)$$

$$\frac{d\theta}{d\tau} = \psi(1-z)^n h(\theta) - \theta \quad (2)$$

where the reaction rate is defined by

$$h(\theta) = \exp\left(\frac{\theta}{1 + \epsilon\theta}\right), \quad (3)$$

and the initial conditions at $\tau = 0$ are

$$z(0) = z^0 = 0, \quad \theta(0) = \theta^0 = 0. \quad (4)$$

All symbols in (1)-(4) are explained in the text and in the Notations.

Parametric sensitivity is concerned with the dependence of system behavior on heat release and heat loss parameters. The problem is very simple if reactant consumption is neglected, i.e., we drop eq. (1) and set $z(t) = 0$ in (2)-(4). Depending on the value of the Semenov parameter ψ , the temperature then either rises to a maximum and subsequently falls back to the ambient (subcritical behavior), or it increases

monotonically and becomes unbounded in finite time (supercritical behavior). The system is stable in the first case and is unstable in the second. The clear distinction between subcritical and supercritical trajectories disappears when reactant consumption is taken into account, because after attaining its maximum θ^* the temperature always returns to the ambient, which is the unique and stable steady state. Nevertheless, there is a characteristic value ψ_c of the Semenov parameter at which the trajectories become very sensitive to variations in parameters and initial conditions. This concept of runaway, also called parametric sensitivity, has been introduced by Bilous and Amundson (1956) in the context of chemical reactor theory. They calculated the sensitivity of the temperature with respect to several input variables along the trajectory corresponding to nominal operating conditions. The system was said to exhibit parametric sensitivity if these sensitivity functions increased to very large values.

In order to eliminate the unspecified threshold on the sensitivities, Thomas and Bowes (1961) and Adler and Enig (1964) proposed criteria for parametric sensitivity based on the occurrence of a positive second-order derivative before the maximum, in the temperature-time and temperature-conversion planes, respectively. These definitions do not require the use of arbitrary threshold values, but their relationship to the original formulation of Bilous and Amundson (1956) is not straightforward. The sensitivity concept was reintroduced by Boddington et al (1983) into the runaway theory. In their formulation the sensitivity of the maximum temperature θ^* with respect to the Semenov number ψ takes its maximum at the critical value ψ_c of ψ . This condition was generalized by Morbidelli and Varma (1988) who noticed that to define the critical Semenov number ψ_c one can use the derivative of θ^* with respect to any parameter p_j instead of $\partial\theta^*/\partial\psi$, since all sensitivities as functions of ψ have their maxima at the same point. The generalized criterion is firmly based on sensitivity concepts and emphasizes that at criticality the maximum temperature θ^* becomes simultaneously

sensitive to small changes in any of the model parameters. The criterion, originally proposed for the explosion model (1)-(4), has been extended to further systems (Morbiddelli and Varma; 1989).

For a general treatment of scaling and self-similarity it is convenient to consider a model of the form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, \mathbf{p}), \quad \mathbf{y}(0) = \mathbf{y}_0, \quad (5)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ and $\mathbf{p} = (p_1, p_2, \dots, p_q)^T$ denote the variables and parameters of the model, respectively. Rabitz and Smooke (1988) observed that the derivatives $\partial y_i / \partial p_j$ and $\partial y_i / \partial t$ frequently satisfy the scaling relations of the form

$$\frac{\partial y_i / \partial p_k}{\partial y_j / \partial p_k} \approx \frac{\partial y_i / \partial t}{\partial y_j / \partial t} \quad (6)$$

for all t . Relations (6) immediately imply that

$$\frac{\partial y_i / \partial p_k}{\partial y_i / \partial p_l} \approx \frac{\partial y_j / \partial p_k}{\partial y_j / \partial p_l}, \quad (7)$$

thus the ratio of sensitivity functions with respect to parameters p_k and p_l is the same for any variable of the model. The self-similarity relations formulated by Rabitz and Smooke (1988) go a step further and show that these ratios are of the form

$$\frac{\partial y_i / \partial p_k}{\partial y_i / \partial p_l} \approx \frac{\sigma_k}{\sigma_l}, \quad (8)$$

for all t , where σ_k and σ_l are constant coefficients. Equation (8) states that the sensitivity functions of a given variable, with respect to a sequence of parameters, will be described by a self-similar set of curves as functions of time, all related by the constants in the vector $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_q)$. Scaling and self-similarity relations have been verified also in steady-state problems such as in steady premixed laminar flames (Vajda et al.; 1990).

Two observations suggest that there exist relations between thermal runaway and self-similarity. First, the sensitivity coefficients $\partial \theta^* / \partial p_j$ of the temperature maximum

θ^* with respect to various parameters p_j ; not only have their extrema at the same value ψ_c , but are also similar as functions of the Semenov number ψ (see Figures 4 and 7 in Morbidelli and Varma; 1988). Second, as we show further in this paper, the sensitivity functions of the model (1)-(4) satisfy the self-similarity relations if and only if the system exhibits critical or supercritical behavior. A further motivation for a joint analysis of the two phenomena is that both seem to be somewhat beyond the scope of usual sensitivity studies. In fact, the main goal of sensitivity analysis is to quantify the influence of individual parameters on system behavior. Thermal runaway and self-similarity are, however, phenomena that apart from constant scaling factors do not depend on the choice of the particular parameter being perturbed.

As shown by Rabitz and Smooke (1988), both scaling and self-similarity conditions for system (5) can be derived by assuming the existence of a dominant independent variable y_n and appropriate functions F_1, \dots, F_{n-1} such that all the other variables y_1, \dots, y_{n-1} can be expressed as

$$y_i(t, p) = F_i(y_n(t, p)), \quad i = 1, \dots, n-1. \quad (9)$$

Notice that the functions F_i explicitly depend neither on time nor on the parameters. The explosion system (1)-(4), however, satisfies self-similarity relations (8), whereas no scaling relations of form (6) were observed. In the present paper self-similarity is derived without assuming (9). Nevertheless, the temperature is shown to be the dominant variable under critical or supercritical conditions. Furthermore, we identify a relationship that can be regarded as a generalization of (9).

2. A NEW GENERALIZED CONDITION FOR PARAMETRIC SENSITIVITY

This section relies on the results of Morbidelli and Varma (1988) who showed that the critical value ψ_c of the Semenov parameter satisfies the relations $|\partial\theta^*(\psi_c)/\partial p_j| \geq$

$|\partial\theta^*(\psi)/\partial p_j|$ for all ψ , where p_j can be any of the parameters in eqs. (1)-(4). The equations will be written in the form

$$\frac{dz}{dt} = \frac{1}{B}(1-z)^n\phi(T) \quad (10)$$

$$\frac{dT}{dt} = \epsilon(1-z)^n\phi(T) - \frac{1}{\epsilon\psi}(T-1), \quad (11)$$

where the temperature dependence of reaction rate constant is given by

$$\phi(T) = \exp\left(\frac{T-1}{\epsilon T}\right), \quad (12)$$

and $t = \psi\tau$. The initial conditions are

$$z(0) = z_0 = 0, \quad T(0) = T_0 = 1. \quad (13)$$

Eq. (12) explicitly shows the role of the activation energy parameter ϵ . The model now has four parameters ψ, B, ϵ , and n . Similarly to Morbidelli and Varma (1988) we consider only $n = 1$ in the calculations. For simplicity the vector notation $\mathbf{y} = (z, T)^T$ will also be used, thereby reducing (10)-(11) to the general form (5).

Using either sensitivity or stability concepts in the analysis of thermal runaway it is natural to study the behavior of system (10)-(13) in the vicinity of the nominal trajectory $\mathbf{y}(t)$, i.e., of the trajectory that corresponds to nominal parameters. A simple approach to this local analysis involves the linear perturbation equation

$$\frac{d}{dt}\delta\mathbf{y} = \mathbf{A}(\mathbf{y})\delta\mathbf{y}, \quad \delta\mathbf{y}(0) = \delta\mathbf{y}^0, \quad (14)$$

where the elements $a_{ij} = \partial F_i / \partial y_j$ of the Jacobian matrix \mathbf{A} are given by

$$a_{11} = -\frac{n}{B}(1-z)^{n-1}\phi(T) \quad (15a)$$

$$a_{12} = \frac{1}{B}(1-z)^n \frac{\partial f}{\partial T}(T) \quad (15b)$$

$$a_{21} = -\epsilon n(1-z)^{n-1}\phi(T) \quad (15c)$$

$$a_{22} = \epsilon(1-z)^n \frac{\partial f}{\partial T}(T) - \frac{1}{\epsilon\psi}, \quad (15d)$$

and

$$\frac{\partial f}{\partial T}(T) = \frac{\phi(T)}{\epsilon T^2}. \quad (15e)$$

Consider first a two-dimensional linear dynamical system of the form (14) but with a constant coefficient matrix A . Based on the text by Hirsch and Smale (1974), Figure 1 summarizes the geometric information on the form of behavior that can be deduced from the characteristic equation

$$\lambda^2 - (tr A)\lambda + det A = 0. \quad (16)$$

The regions corresponding to different forms of behavior are divided by the parabole $\Delta = 0$, where $\Delta = (tr A)^2 - 4det A$ is the discriminant of the quadratic equation (16). Regions I through IV correspond to stable nodes, stable spirals, unstable spirals, and unstable nodes, respectively. The region with $det A < 0$, not shown in Figure 1, corresponds to saddle behavior.

Since the coefficient matrix in (14) is not constant, the characteristic of the local linearization changes as the point $y(t)$ moves along the nominal trajectory. The determinant of A is given by

$$det A(y) = \frac{n}{B\epsilon\psi} (1-z)^{n-1} \phi(T). \quad (17)$$

Since $0 \leq z \leq 1$, $det A \geq 0$ for all t , and the linear approximation (14) never exhibits saddle-type behavior. Furthermore,

$$tr A(y) = (1-z)^{n-1} \phi(T) \left(\frac{1-z}{T^2} - \frac{n}{B} \right) - \frac{1}{\epsilon\psi}. \quad (18)$$

If the initial conditions are $z^0 = 0$ and $T^0 = 1$, then at $t = 0$

$$tr A(y^0) = 1 - \frac{n}{B} - \frac{1}{\epsilon\psi}, \quad det A(y^0) = \frac{n}{B\epsilon\psi}, \quad (19)$$

whereas at $t \rightarrow \infty$ we have $\mathbf{y}^\infty = (1, 1)^T$, and

$$\text{tr} \mathbf{A}(\mathbf{y}^\infty) = -\frac{1}{\epsilon \psi}, \quad \det \mathbf{A}(\mathbf{y}^\infty) = 0. \quad (20)$$

Thus, at $t = 0$ the local behavior of the system is a sink if $\psi \leq B/[\epsilon(B - n)]$, and it always becomes a sink when $t \rightarrow \infty$. In these regions a local perturbation exponentially decays to the nominal trajectory, and no thermal runaway is possible unless the system locally behaves as a source on some time interval. Indeed, apart from very small values of the Semenov number, we have $\text{tr} \mathbf{A} > 0$ along some segments of the trajectory. Trajectories corresponding to $n = 1$, $B = 50$, $\epsilon = 0.1$, and three different values of ψ are shown in Figure 1. Each point in this plane describes the geometric character of the perturbation equation (14) around the point $\mathbf{y}(t)$. This character changes as $\mathbf{y}(t)$ moves along the nominal trajectory, and according to Figure 1 it goes through the following stages: stable node, stable spiral, unstable spiral, unstable node, and then backward all the way to the stable node.

The geometric definition of thermal runaway due to Adler and Enig (1964) and both sensitivity-based definitions by Boddington et al. (1983) and by Morbidelli and Varma (1988) consider the behavior near or at the temperature maximum T^* . Therefore we also consider the point $\mathbf{y}(t^*)$, where t^* denotes the time of the temperature maximum. Instead of looking for a positive second-order derivative before t^* (Adler and Enig, 1964) or for the maximum of the sensitivity $\partial T^*/\partial p_j$ as a function of ψ (Morbidelli and Varma, 1988), we ask how a perturbation $\delta \mathbf{y}(t^*)$ applied at time t^* will propagate when $t > t^*$. There are two forms of this behavior. The equilibrium point $\delta \mathbf{y} = 0$ of the perturbation equations (14) is either stable and then the system returns to the nominal trajectory, or the linear approximation is unstable and then the perturbation $\delta \mathbf{y}(t^*)$ is amplified on some interval $[t^*, t]$. It is reasonable to identify

criticality with the conditions leading to the maximum of such amplification. Consider a small time step $\delta t = t - t^*$, then the solution of (14) is approximated by

$$\delta \mathbf{y}(t) = \exp[\mathbf{A}(\mathbf{y}^*)\delta t]\delta \mathbf{y}(t^*). \quad (21)$$

By the definition of the matrix norm

$$\max \frac{\|\delta \mathbf{y}(t)\|}{\|\delta \mathbf{y}(t^*)\|} = \exp[Re(\lambda_{\max})\delta t], \quad (22)$$

where $\|\mathbf{y}(t)\|$ denotes the Euclidean norm of the vector $\mathbf{y}(t)$, and $Re(\lambda_{\max})$ is the largest real part of the two eigenvalues of $\mathbf{A}(\mathbf{y})$ at $t = t^*$.

We define criticality as the point in the parameter space at which $Re(\lambda_{\max})$ takes its maximum, provided this maximum is nonnegative. Therefore, criticality implies maximum sensitivity of the nominal trajectory to perturbations applied at time t^* . If $Re(\lambda_{\max}) < 0$, then by (21) all perturbations decay and no runaway is possible. If we consider one of the model parameters, say the Semenov number ψ , and keep all the others fixed, then the critical value ψ_c is that value of ψ which maximizes $Re(\lambda_{\max})$ at $\mathbf{y}(t^*)$. As shown in Figure 2 for $\epsilon = 0.1$, keeping the other parameters fixed $Re(\lambda_{\max})$ exhibits a maximum at a specific value of ψ which is then defined as the critical Semenov number ψ_c .

In Table 1 we compare the critical value ψ_c given by our definition with the values derived by Adler and Enig (1964) and Morbidelli and Varma (1988). Notice that the latter condition is based on the use of the sensitivity coefficients $\partial T^*/\partial p_j$, where p_j is one of the model parameters, and the value of ψ_c may depend on the choice of p_j . Therefore Table 1 lists the smallest and largest values found in this way. For each value of B we also show the temperature maximum T^* , the discriminant Δ and the value of $Re(\lambda_{\max})$ at T^* . For $B \geq 30$ the agreement with the previous criteria is very good. For smaller B 's the ψ_c values predicted by the three criteria start to deviate, with our criterion resulting in the lowest estimates of ψ_c . According to Table

2 similar conclusions hold for $\epsilon = 0$. Notice that in Table 2 we list the values of the maximum dimensionless temperature θ^* used in eqs. (1)-(4) instead of the maximum temperature T^* . This renders our results directly comparable to those of Morbidelli and Varma (1988).

Since the new condition is based on local linearization and eigenvalue analysis, it is related to the work of Gray and Sherrington (1972a, 1972b) who derived critical values for the Semenov number on the base of Liapunov's stability theorems. Gray and Sherrington (1972b) noticed that, compared to their criticality condition, the method of Adler and Enig (1964) always overestimates the stable region. This overestimation is seen experimentally since the predicted stable temperatures are higher than observed in practice. According to Tables 1 and 2 our criterion also gives lower critical values than predicted by Adler and Enig (1964).

For small B 's the previous criteria not only overpredict the stable region but the predictions become rather unreliable. This is not surprising since with decreasing values of B and ϵ^{-1} the magnitude of the temperature maximum itself becomes rather small, and the system gradually loses its sensitivity potential. While this non-explosive region is physically not very interesting, our criterion shows that for certain regions of the parameters n , B , and ϵ we have $Re(\lambda_{max}) < 0$ for all values of ψ , and there exists no critical Semenov number. None of the previous criteria can give such a clear result. As noticed by Morbidelli and Varma (1988), all criteria based on the topology of the temperature-conversion or temperature-time profiles *a priori* assume the existence of a critical point at any value of B . This criticism is valid, but the situation is not much better with the criterion of Morbidelli and Varma (1988). In fact, the only sign that indicates the uncertainty in the value of ψ_c , or possibly the nonexistence of runaway, is the deviation among predictions based on the choice of different parameters p_j in the generalized condition. According to our criterion (Table 1), for $B=7$ the local linearization is asymptotically stable for any value of ψ , and any perturbation of the

nominal trajectory applied at time t^* decays to zero. Thus, there exists no runaway at these parameters. Similarly, there is no runaway at $\epsilon = 0$ if $B \leq 5$. As shown in Figure 3, decreasing the value of B at any fixed ϵ we reach a lower bound B_l such that no critical Semenov number exists for $B < B_l$. This explains why predictions by other methods can deviate from each other even by orders of magnitude in this region of the parameters.

We conclude this section with some remarks on the form of trajectories in the plane defined by the coordinates trA and $detA$ as shown in Figure 1. By (17) for $n = 1$ we have $detA = \phi(T)/B\psi^0$. Since ϕ defined by (12) is a monotonically increasing function of T , for $n = 1$ the maximum of each curve in Figure 1 (i.e., the maximum of $detA$) is at the temperature maximum T^* . Using our generalized criterion we increase ψ while keeping the other parameters fixed and observe when $Re(\lambda_{max})$ reaches its maximum. As shown in Figure 1, this maximum first moves to the right with increasing ψ , then moves to the left when ψ passes the critical value ψ_c . If the maximum is in region III, then $Re(\lambda_{max}) = trA/2$, and thus it can be identified with the utmost right position of the maximum. In region IV $Re(\lambda_{max}) = (trA + \sqrt{\Delta})/2$, and the geometric interpretation is not so simple. On the boundary of the two regions $\Delta = 0$ and the Jacobian matrix has a pair of identical real eigenvalues. According to Tables 1 and 2, Δ has small negative values at criticality in most cases. The maximum of the curve is then in region III close to the boundary, and at criticality the system behaves as an unstable spiral with small imaginary components in the eigenvalues. This behavior, however, suddenly changes at sufficiently high values of the heat of reaction parameter B . For example, with $\epsilon = 0$ and $n = 1$ such a change occurs at $B \geq 26$. Then the discriminant Δ becomes large and positive around the critical point, and thus the maximum is in region IV. The criticality is very sharp: a slight increase in ψ will move the maximum of the curve far to the left into region I.

3. EFFECTS OF CRITICAL CONDITIONS ON SELF-SIMILARITY

Figures 4 through 9 demonstrate self-similarity and its relation to criticality for model (10)-(13) by presenting the sensitivity functions with respect to the parameters $p_1 = \psi$, $p_2 = B$, and $p_3 = \epsilon$ calculated at $n = 1$, $B = 30$, $\epsilon = 0.1$, and different values of ψ . Figures 4 and 5 were obtained at $\psi = 0.55$ that generates subcritical behavior. The sensitivities with respect to B and ϵ are small compared to the ones with respect to ψ , and not much similarity is seen among the three functions. There is, however, noticeable similarity at the critical value $\psi_c = 0.6107$ (Figures 6 and 7). As expected from the definition of parametric sensitivity, at $\psi = \psi_c$ the maximum of the function $\partial T / \partial p_j$ occurs at $t = t^*$, the time of temperature maximum ($t^* = 4.81$ for the conditions shown in Figure 6). A slight increase in the value of ψ moves the system into the supercritical region and alters the form of the temperature sensitivity function which now changes sign close to t^* ($t^* = 4.27$ for the conditions of Figure 8). According to Figures 8 and 9 the similarity of sensitivity function is preserved in the supercritical region.

By (8) self-similarity assumes the existence of constants a_j such that $\partial y_k / \partial p_j \approx a_j (\partial y_k / \partial p_1)$ on some time interval for $k = 1, 2$ and $j = 2, 3$. Since these relations are approximate, we introduce the sum of squares objective functions

$$\bar{Q}_k(a_2, a_3) = \sum_{j=2}^3 \sum_{i=1}^m \left[\frac{\partial \tilde{y}_k(t_i)}{\partial p_j} - a_j \frac{\partial \tilde{y}_k(t_i)}{\partial p_1} \right]^2, \quad (23)$$

where $y_1 = z$, $y_2 = T$, m is the number of selected time points t_1, \dots, t_m , and

$$\frac{\partial \tilde{y}_k(t_i)}{\partial p_j} = \frac{\partial y_k(t_i) / \partial p_j}{(\partial y_k / \partial p_j)_{\max}} \quad (24)$$

is the normalized sensitivity function with $(\partial y_k / \partial p_j)_{\max}$ representing the maximum sensitivity. We use this particular normalization to give approximately equal weights

to the two sensitivity functions in (23). Following a least squares estimation of the factors a_2 and a_3 , the degree of similarity is measured by the residual sums of squares

$$Q_k = \min \bar{Q}_k(a_2, a_3), \quad k = 1, 2. \quad (25)$$

Figures 10 and 11 show how Q_1 and Q_2 depend on the Semenov parameter at $B = 30$ and $B = 50$, respectively. The plots were generated at $n = 1$ and $\epsilon = 0.1$, selecting 50 equidistant points with time steps $\Delta t = t_{i+1} - t_i = 0.2$. The residual sum (25) quickly decreases as ψ approaches its critical value ($\psi_c = 0.6107$ and $\psi_c = 0.533$ in Figures 10 and 11, respectively), and remains almost constant in the supercritical region. A slight local minimum can be observed close to the point of criticality. In critical and supercritical regions the residual errors defined by $s_k = \sqrt{Q_k/(2m - 2)}$ are $s_k \approx 0.045$ and $s_k \approx 0.01$ for $B = 30$ and $B = 50$, respectively. This shows high degrees of similarity in both cases. Notice that similarity improves with increasing values of B when considering critical or supercritical points.

In this section we will show that self-similarity follows from two properties of the simple explosion system described by eqs. (10) and (11). The first property is strong coupling of the conversion variable to the temperature. The second is the pseudohomogeneous behavior of the corresponding sensitivity equations on an open neighborhood of t^* , the point of maximum temperature. These properties will be discussed in turn.

3A. STRONG COUPLING APPROXIMATION

For notational simplicity we write equations (10) and (11) in the general form as

$$\frac{dz}{dt} = f_1(z, T, p) \quad (26a)$$

$$\frac{dT}{dt} = f_2(z, T, p). \quad (26b)$$

Differentiating (26) with respect to the parameter p_j one obtains the sensitivity equations

$$\frac{d}{dt} \frac{\partial z}{\partial p_j} = \frac{\partial f_1}{\partial z} \frac{\partial z}{\partial p_j} + \frac{\partial f_1}{\partial T} \frac{\partial T}{\partial p_j} + \frac{\partial f_1}{\partial p_j} \quad (27a)$$

$$\frac{d}{dt} \frac{\partial T}{\partial p_j} = \frac{\partial f_2}{\partial z} \frac{\partial z}{\partial p_j} + \frac{\partial f_2}{\partial T} \frac{\partial T}{\partial p_j} + \frac{\partial f_2}{\partial p_j}. \quad (27b)$$

To study the coupling of the two variables we decouple them by assuming that $\partial T / \partial p_j$ is a known function and considering the first equation (27a) separately. This equation can be solved through the Green's function $g_1(t, t')$ which is the solution of the time-variable linear differential equation

$$\frac{d}{dt} g_1(t, t') = \frac{\partial f_1}{\partial z}(t) g_1(t, t') + \delta(t - t'), \quad (28)$$

where $\delta(t - t')$ denotes the Dirac impulse function, and the initial conditions are given by $g_1(t', t') = 1$, and $g_1(t, t') = 0$ for $t < t'$. The solution of (28) is given by

$$g_1(t, t') = \exp \int_{t'}^t \frac{\partial f_1}{\partial z}(\tau) d\tau, \quad (29)$$

and in terms of $g_1(t, t')$ the conversion sensitivity functions are

$$\frac{\partial z}{\partial p_j}(t) = \int_0^t g_1(t, t') \frac{\partial f_1}{\partial T}(t') \frac{\partial T}{\partial p_j}(t') dt' + \int_0^t g_1(t, t') \frac{\partial f_1}{\partial p_j}(t') dt'. \quad (30)$$

Let us now decouple the variables by fixing the temperature at its nominal profile, i. e., considering $T(t)$ as an external variable, independent of the system parameters. This is equivalent to the condition $\partial T / \partial p_j = 0$ for all t , and the conversion sensitivity function is reduced to the second term in eq. (30). This term, defined as

$$[\frac{\partial z}{\partial p_j}(t)]_T = \int_0^t g_1(t, t') \frac{\partial f_1}{\partial p_j}(t') dt' \quad (31)$$

is called the constrained temperature sensitivity function of the conversion. It is the sensitivity function corresponding to a process in which the temperature of the reaction vessel is controlled to exactly follow a prescribed profile in spite of the perturbations

in the system parameters. Equation (30) is a decomposition of the sensitivity function where the constrained temperature sensitivity term (31) measures the direct effects of parameter perturbations on the conversion, whereas the first term corresponds to the indirect effects (i. e., the parameter perturbations that change the temperature which, in turn, affects the conversion by altering the reaction rate).

Figure 12 shows the constrained temperature sensitivity functions of the conversion at $n = 1$, $B = 30$, $\epsilon = 0.1$, and the critical value $\psi_c = 0.6107$ of the Semenov number. We compare these functions to the original conversion sensitivity coefficients shown in Figure 7. While all functions are small at the beginning and also for large values of t (beyond the interval shown in the Figures), there exists a characteristic window $[t_1, t_2]$ on which the first term in (30) is much larger than the second. The values of t_1 and t_2 are not unique. For example, selecting any $t_1 > 4.5$ and $t_2 < 10$ on $[t_1, t_2]$ each sensitivity coefficient in Figure 7 is at least five times larger than the corresponding constrained sensitivity in Figure 12. Retaining only the dominant term in (30) leads to the approximation

$$\frac{\partial z}{\partial p_j}(t) \approx \int_0^t g_1(t, t') \frac{\partial f_1}{\partial T}(t') \frac{\partial T}{\partial p_j}(t') dt' \quad (32)$$

for $t_1 < t \leq t_2$. We refer to (32) as the strong coupling approximation, since it is based on the strong coupling of the conversion variable to the temperature which implies that on $[t_1, t_2]$ any parameter perturbation dominantly affects the conversion through the induced perturbation in the temperature of the reaction vessel.

Approximation (32) helps to understand how conversion and temperature sensitivity functions are related to each other. Due to (15b) $\partial f_1 / \partial T = a_{12} \geq 0$ for all t , and by (29) $g_1(t, t') > 0$ for all t and t' . Thus the sign of the integrand in (32) is determined by the sign of $\partial T / \partial p_j$. As shown in Figure 6, for $\psi = 0.6107$ the functions $\partial T / \partial p_j$ change sign around $t = 6$, and then remain small. Accordingly, the conversion sensitivities in Figure 7 slowly decrease when $t > 6$. For $\psi = 0.63$ there is a sign

change in the temperature sensitivities at $t \approx 4.25$, but their magnitudes become large again (Figure 8). This explains why the conversion sensitivities shown on Figure 9 and approximated by the integral (32) quickly decrease when $t > 4.25$.

The strong coupling approximation is clearly related to self-similarity, although does not completely explain it. In particular, due to (32) the self-similarity of the temperature sensitivity functions $\partial T/\partial p_j$, $j = 1, 2, 3$, implies the self-similarity of the conversion sensitivity functions $\partial z/\partial p_j$, $j = 1, 2, 3$. To show this relationship assume that $\partial T/\partial p_i$ and $\partial T/\partial p_j$ are self-similar over the interval $[t_1, t_2]$, i. e., there exists a constant a_i such that $\partial T(t)/\partial p_j \approx a_i \partial T(t)/\partial p_i$ for all $t_1 < t \leq t_2$. It immediately follows from the linearity of the integral operator in (32) that $\partial z(t)/\partial p_j \approx a_i \partial z(t)/\partial p_i$ for $t_1 \leq t \leq t_2$.

The validity of (32), in turn, is related to parametric sensitivity. As discussed in Section 2, Morbidelli and Varma (1988) showed that the sensitivity coefficients $\partial T^*/\partial p_j$ as functions of ψ have very sharp maxima at the critical Semenov number ψ_c . When the Semenov number approaches its critical value, the first term in (30) becomes more and more dominant. Therefore, if there exist any parameter value such that the strong coupling approximation (32) applies to a model, then it certainly applies close to the point of criticality. Although in the supercritical region the maximum of $\partial T/\partial p_j$ precedes t^* , according to our calculations the first term in (30) remains dominant on an interval $[t_1, t_2]$.

Similarly to the constrained temperature sensitivities of the conversion we can calculate the constrained conversion sensitivities of the temperature. In terms of the Green's function

$$g_2(t, t') = \exp \int_{t'}^t \frac{\partial f_2}{\partial T}(\tau) d\tau, \quad (33)$$

the solution of the temperature sensitivity equation (27b) is given by

$$\frac{\partial T}{\partial p_j}(t) = \int_0^t g_2(t, t') \frac{\partial j_2}{\partial z}(t') \frac{\partial z}{\partial p_j}(t') dt' + \int_0^t g_2(t, t') \frac{\partial f_2}{\partial p_j}(t') dt', \quad (34)$$

where the second term

$$\left[\frac{\partial T}{\partial p_j}(t)\right]_z = \int_0^t g_2(t, t') \frac{\partial f_2}{\partial p_j}(t') dt' \quad (35)$$

is defined as the constrained conversion sensitivity function of the temperature. The symmetry, however, ends at this point. Comparing Figures 8 and 13 shows that on some time interval containing t^* the constrained conversion sensitivity functions (35) of the temperature are even larger than the corresponding full sensitivity functions (34). The different behavior of the two variables will be further discussed in the next section.

3B. PSEUDOHOMOGENEOUS SENSITIVITY EQUATIONS

The sensitivity equations (27) are inhomogeneous due to the terms $\partial f_1/\partial p_j$ and $\partial f_2/\partial p_j$. It follows, however, from the strong coupling approximation (32) that the term $\partial f_1/\partial p_j$ in (27) can be neglected. Operating with $(\partial/\partial t - \partial f_1/\partial z)$ on eq. (32) and using eq. (28) yields

$$\frac{d}{dt} \frac{\partial z}{\partial p_j}(t) = \frac{\partial f_1}{\partial z}(t) \frac{\partial z}{\partial p_j}(t) + \frac{\partial f_1}{\partial T}(t) \frac{\partial T}{\partial p_j}(t). \quad (36)$$

Comparing (36) to (27a) implies that

$$\partial f_1(t)/\partial p_j \approx 0 \quad (37)$$

for $t_1 < t < t_2$.

As discussed, the strong coupling approximation does not apply to the temperature sensitivity equation. Therefore, it is an independent observation that the direct term $\partial f_2/\partial p_j$ is nevertheless small in (27b) over some interval $[t_1, t_2]$. For example, Figure 14 shows the four terms $(\partial f_1/\partial z)(\partial z/\partial \psi) + (\partial f_1/\partial T)(\partial T/\partial \psi)$, $(\partial f_2/\partial z)(\partial z/\partial \psi) + (\partial f_2/\partial T)(\partial T/\partial \psi)$, $\partial f_1/\partial \psi$, and $\partial f_2/\partial \psi$ separately at the critical point corresponding

to $B = 30$. In this particular case $\partial f_1/\partial \psi = 0$ for all t , but the magnitude of $\partial f_2/\partial \psi$ is also relatively small on the interval $3 < t < 6.5$. This suggests the approximation

$$\partial f_2(t)/\partial p_j \approx 0 \quad (38)$$

for $t_1 < t < t_2$. The approximation (38) may seem contradictory with the observation that (35) is not small, but this argument neglects the behavior of $g_2(t, t')$ as will be discussed below. Notice that by intuition the derivatives $\partial f_i/\partial p_j$ are also expected to be relatively small after an induction period. Indeed, an explosion or flame is difficult to stop once started, thus the process must be rather insensitive to perturbations such as a change in the ambient temperature.

First we show why the strong coupling approximation of the form (32) applies to the conversion. Evaluating the integral (31) on the subintervals $[0, t_1]$ and $[t_1, t']$ separately, i.e., in the form

$$\left[\frac{\partial z}{\partial p_j}(t)\right]_T = \int_0^{t_1} g_1(t, t') \frac{\partial f_1}{\partial p_j}(t') dt' + \int_{t_1}^t g_1(t, t') \frac{\partial f_1}{\partial p_j}(t') dt', \quad (39)$$

neglecting the second term due to (37), and rewriting the first term using the relation $g_1(t, t') = g_1(t, t_1)g_1(t_1, t')$ we have:

$$\left[\frac{\partial z}{\partial p_j}(t)\right]_T \approx g_1(t, t_1) \int_0^{t_1} g_1(t_1, t') \frac{\partial f_1}{\partial p_j}(t') dt' = g_1(t, t_1) \left[\frac{\partial z}{\partial p_j}(t_1)\right]_T. \quad (40)$$

By (15a) $\partial f_1/\partial z \leq 0$ for all t , and by (26) $g_1(t, t')$ is a quickly decreasing function of t for any $t > t_1$. Since neither (32) nor (37) apply for $t < t_1$, the constrained sensitivity is not necessarily small on this interval, but will quickly diminish for $t > t_1$, and the first term in (30) becomes dominant.

To prove that the constrained sensitivity function $[\partial T/\partial p_j]_z$ can be relatively large in spite of assumption (38), we now evaluate the integral in (35) on the subintervals $[0, t_1]$ and $[t_1, t']$ separately. Then, similarly to (40), the assumption (38) implies that

$$\left[\frac{\partial T}{\partial p_j}(t)\right]_z \approx g_2(t, t_1) \left[\frac{\partial T}{\partial p_j}(t_1)\right]_z. \quad (41)$$

The functions $g_1(t, t')$ and $g_2(t, t')$ are, however, very different. By (15d) there exists a time interval such that $\partial f_2 / \partial T > 0$, since a positive feedback through the temperature over some period of time is a trivial necessary condition for explosion. Therefore, while $g_1(t, t')$ quickly diminishes for $t > t'$, by (33) $g_2(t, t')$ increases almost exponentially on the interval with $\partial f_2 / \partial T > 0$. By (41) $[\partial T / \partial p_j]_z$ also increases during this period of time following t_1 , and the constrained sensitivity function can be large in spite of (38).

Eq. (36) has a further implication. Since f_1 does not explicitly depend on the parameters for $t_1 < t < t_2$, (27a) reduces to

$$\frac{dz}{dt} = f_1(z, T) \quad (42)$$

over the same interval. Let $T_{[t_1, t]}(p)$ denote the segment of the temperature profile at parameters p over the time interval $[t_1, t]$, and consider this function as an input to (42). Assuming that the conversion $z(t_1, p)$ at time t_1 is small, the solution of (42) for some interval $t_1 + \delta < t < t_2$ is of the form

$$z(t, p) = \Psi(T_{[t_1, t]}(p)), \quad (43)$$

where Ψ is a functional, and $\delta > 0$ is a positive constant such that the term $z(t_1, p)$ can be neglected for $t > t_1 + \delta$. Since the functional Ψ does explicitly depend neither on time nor the parameters, (43) is a generalization of the relation (9) with the temperature as dominant dependent variable. The conversion z at time t depends, however, on an entire segment $T_{[t_1, t]}$ of the temperature profile and not only on the actual temperature at time t .

3.C. THE ORIGIN OF SELF-SIMILARITY

This subsection shows that the validity of the strong coupling approximation (32) and the pseudohomogeneity assumption (38) together imply self-similarity. Let $G(t, t')$

denote the 2×2 Green's function matrix for the equations (10)-(11). Then $\mathbf{G}(t, t')$ is the solution of the matrix differential equation

$$\frac{d}{dt} \mathbf{G}(t, t') = \mathbf{A}(t) \mathbf{G}(t, t') + \delta(t - t') \mathbf{I}, \quad (44)$$

where the elements of $\mathbf{A}(t)$ are given by (15) at $\mathbf{y}(t)$, and the initial conditions are $\mathbf{G}(t', t') = \mathbf{I}$, and $\mathbf{G}(t, t') = 0$ for $t < t'$. In terms of $\mathbf{G}(t, t')$ the sensitivity functions are

$$\begin{pmatrix} \partial z(t)/\partial p_j \\ \partial T(t)/\partial p_j \end{pmatrix} = \int_0^t \mathbf{G}(t, t') \begin{pmatrix} \partial f_1(t')/\partial p_j \\ \partial f_2(t')/\partial p_j \end{pmatrix} dt'. \quad (45)$$

We evaluate the integral in (45) over the intervals $[0, t_1]$ and $[t_1, t]$ separately. By (37) and (38) the integrals on $[t_1, t]$ vanish. Exploiting the relationship $\mathbf{G}(t, t') = \mathbf{G}(t, t_1) \mathbf{G}(t_1, t')$, eq. (45) is reduced to

$$\begin{pmatrix} \partial z(t)/\partial p_j \\ \partial T(t)/\partial p_j \end{pmatrix} = \mathbf{G}(t, t_1) \begin{pmatrix} \partial z(t_1)/\partial p_j \\ \partial T(t_1)/\partial p_j \end{pmatrix} \quad (46)$$

on the interval $t_1 < t < t_2$. The first equation of (46) is

$$\frac{\partial z}{\partial p_j}(t) = g_{11}(t, t_1) \frac{\partial z}{\partial p_j}(t_1) + g_{12}(t, t_1) \frac{\partial T}{\partial p_j}(t_1), \quad (47)$$

where g_{11} and g_{12} are the two entries in the first row of \mathbf{G} . By the strong coupling approximation (32), if $\partial T/\partial p_j = 0$ then this implies $\partial z/\partial p_j \approx 0$. This is possible if and only if the second term in (47) is much larger than the first one, leading to the approximation

$$\frac{\partial z}{\partial p_j}(t) \approx g_{12}(t, t_1) \frac{\partial T}{\partial p_j}(t_1). \quad (48)$$

Since $g_{11}(t_1, t_1) = 1$ and $g_{12}(t_1, t_1) = 0$, (48) can be valid only for $t_1 + \delta < t < t_2$, where δ is a positive constant. To calculate the Green's function matrix $\mathbf{G}(t, t_1)$ we used the relationship $\mathbf{G}(t, t_1) = \mathbf{G}(t, 0) \mathbf{G}^{-1}(t_1, 0)$, where the last two matrices were obtained by solving equation (44) with $t' = 0$. According to these calculations, in critical and supercritical regions there exist t_1 and t_2 such that the Green's function

g_{12} is similar to $\partial z/\partial p_j$ over the interval $[t_1 + \delta, t_2]$. This result supports the validity of the assumption (38). Furthermore, δ can be chosen so small that it will be omitted in the following. The conversion sensitivity itself turns out to be very small outside the interval $[t_1, t_2]$. However, the validity of (48) does not imply that g_{11} is small compared to g_{12} . In fact, the two functions can have comparable magnitudes, and it is the ratio of the sensitivities $\partial z/\partial p_j$ and $\partial T/\partial p_j$ at t_1 that makes the second term in (47) dominant.

Equation (48) implies self-similarity. Indeed,

$$\frac{\partial z(t)/\partial p_j}{\partial z(t)/\partial p_i} \approx \frac{\partial T(t_1)/\partial p_j}{\partial T(t_1)/\partial p_i}, \quad (49)$$

where the right hand side is constant. The choice of t_1 is, however, not unique, and (49) must be valid for any $t'_1 \geq t_1$ that is sufficiently close to t_1 . Thus the right hand side of (49) is the same constant for all $t_1 < t < t_2$, and hence both the conversion and temperature sensitivity function satisfy the self-similarity relations (8).

Some of the relations derived here can be used to explain the origin of scaling relations (6) if present in the system. Differentiating (26) with respect to time yields

$$\frac{d\dot{z}}{dt} = \frac{\partial f_1}{\partial z} \dot{z} + \frac{\partial f_1}{\partial T} \dot{T} \quad (50a)$$

$$\frac{d\dot{T}}{dt} = \frac{\partial f_2}{\partial z} \dot{z} + \frac{\partial f_2}{\partial T} \dot{T}. \quad (50b)$$

This is the homogeneous part of eq. (27), and for $t \geq t_1$ its solution is given by

$$\begin{pmatrix} \dot{z}(t) \\ \dot{T}(t) \end{pmatrix} = \mathbf{G}(t, t_1) \begin{pmatrix} \dot{z}(t_1) \\ \dot{T}(t_1) \end{pmatrix}. \quad (51)$$

The first equation of (51) in a more explicit form is

$$\dot{z}(t) = g_{11}(t, t_1) \dot{z}(t_1) + g_{12}(t, t_1) \dot{T}(t_1). \quad (52)$$

If the second term in (52) dominates, i.e.,

$$\dot{z}(t) \approx g_{12}(t, t_1) \dot{T}(t_1), \quad (53)$$

then

$$\frac{\partial z(t)/\partial p_j}{dz(t)/dt} \approx \frac{\partial T(t_1)/\partial p_j}{dT(t_1)/dt}. \quad (54)$$

Similarly to (49), the value of t_1 in (54) is not unique. Therefore, the right hand side must be the same constant for all $t \geq t_1$, and the scaling relation of the form (6) follows. For the explosion system (10)-(11), however, the first term in (52) is not negligible, and the variables do not satisfy any scaling relations as it can be readily tested by calculations. This result emphasizes that (43) is a generalization of (9), since assuming a relation of the form (9) implies both scaling and self-similarity (Rabitz and Smooke, 1988).

4. CONCLUSIONS

Both thermal runaway and self-similarity are defined in terms of parameter sensitivity functions but are independent of the choice of particular parameters being perturbed. In thermal runaway the critical value of the Semenov number leads to the maximum of the sensitivity $\partial T^*/\partial p_j$, where T^* is the maximum temperature. As shown by Morbidelli and Varma (1988), p_j generally can be any of the model parameters. Many dynamical systems also satisfy self-similarity relations, and the sensitivity functions of each variable with respect to various parameters are identical up to constant scaling factors.

We consider the basic model in thermal explosion theory, i.e., a well-stirred system in which an exothermic, irreversible reaction occurs, and show that thermal runaway implies self-similarity. The analysis proceeds in several steps leading to interesting intermediate results. First, a new generalized condition for thermal runaway is introduced. As is well known, the concept of thermal runaway is not well defined at

low values of the heat-of-reaction parameter, and the critical points predicted by Morbidelli and Varma (1988) start to depend on the actual choice of the parameter p_j used in the condition. Here the critical condition is defined as the point in the parameter space at which the trajectory exhibits maximum sensitivity to arbitrary, unstructured perturbations applied at the temperature maximum. We show that at this point the largest real part of the two eigenvalues of the Jacobian matrix reaches its maximum. If this maximum is negative, then no thermal runaway is possible. None of the known conditions for parametric sensitivity gives such a clear result. Since it is based on local linearization and eigenvalue analysis, our condition emphasizes the dual origin of thermal runaway, rooted both in stability and sensitivity concepts.

Calculations show that the explosion system satisfies self-similarity relations only under critical and supercritical conditions for thermal runaway. At criticality the temperature becomes the dominant variable, and any perturbation in the parameters affects the conversion by altering the temperature and thereby the reaction rate, whereas the direct, quasi-isothermic effects of parameter perturbations on the conversion are negligible. This results in a simple functional dependence between the temperature and conversion sensitivity functions termed here as the strong coupling approximation. In addition to the strong coupling, criticality in the explosion system implies that after an induction period the sensitivity equations are nearly homogeneous, i.e., the direct effects of parameter perturbations applied at this stage of the reaction are negligibly small.

Both the strong coupling approximation and the pseudohomogeneity of sensitivity equations follow from critical or supercritical behavior and can be directly tested by numerical calculations. On the other hand, these two properties together imply self-similarity. Furthermore, the self-similarity among all sensitivity functions and the dominant role of the temperature shows that restricting consideration to the temperature in the definition of thermal runaway (Bowes, 1961; Adler and Enig, 1964;

Boddington et al, 1983; Morbidelli and Varma, 1988) preserves the generality of the concept. These are the main results of the paper.

Since we restrict consideration to a simple system with only two variables, an important question is whether the results can be generalized to more complex systems. Based on some preliminary calculations the answer is positive. In particular, we studied the case of two consecutive reactions in a pseudohomogeneous tubular reactor (Morbidelli and Varma, 1989). It turns out that one of the eigenvalues of the Jacobian matrix for this system has a large negative real part all the time along the trajectories corresponding to nearly critical conditions, whereas the other two eigenvalues exhibit exactly the same behavior as described in Section 2. Self-similarity is also observed if and only if the conditions are critical or supercritical, and the strong coupling approximation applies to both conversion variables. This makes all our results applicable, but details are beyond the scope of the present paper.

ACKNOWLEDGEMENT

The authors wish to thank the Department of Energy and The Air Force Office of Scientific Research for support of this research.

NOTATION

- A** Jacobian matrix of entries a_{ij} defined by (15)
- B** $(-\Delta H)C^i/C_p\rho_f\bar{T}_a\epsilon$, heat of reaction dimensionless parameter
- C** reactant concentration, mol m^{-3}
- C_p** mean specific heat of reactant mixture, $\text{J K}^{-1} \text{kg}^{-1}$
- det A** determinant of matrix **A**
- E** activation energy, J mol^{-1}
- f** right hand side of the vector differential equation (5)
- F_i** scalar function defined by (9)
- G** Green's function matrix, solution of eq. (44)
- g_{ij}** entries of the Green's function matrix **G**
- h(θ)** $\exp[\theta/(1 + \epsilon\theta)]$, temperature dependence of reaction rate constant
- k** reaction rate constant, $\text{mol m}^{-3} \text{s}^{-1}$
- n** reaction order
- p** parameterization vector in eq. (5)
- \bar{Q}_k** sum of squares function defined by (23)
- R** ideal gas constant, $\text{J K}^{-1} \text{mol}^{-1}$
- Re(λ)** real part of the eigenvalue λ of the Jacobian matrix **A** at the temperature maximum T^*
- S_v** external surface area per unit volume, m^{-1}
- T** \bar{T}/\bar{T}_a , dimensionless temperature
- \bar{T}** absolute temperature of reacting mixture, K
- \bar{T}_a** absolute ambient temperature, K
- \bar{t}** time, s
- t** $\tau\psi$, dimensionless time
- trA** trace of the matrix **A**
- U** overall heat transfer coefficient, $\text{W m}^{-2} \text{K}^{-1}$

y dependent variables in the differential equations (5)
 z $(C^i - C)/C^i$, conversion
 β $US_v/C_p\rho_f k(\bar{T}_a)(C^i)^{n-1}$, dimensionless heat transfer parameter
 Δ discriminant of the quadratic equation (16)
 ΔH enthalpy of reaction, J mol⁻¹
 $\delta(t)$ Dirac delta function
 δy perturbation of the nominal trajectory
 ϵ $R\bar{T}_a/E$, dimensionless activation energy parameter
 θ $(\bar{T} - \bar{T}_a)/\bar{T}_a\epsilon$, dimensionless temperature
 λ_{max} eigenvalue of the Jacobian matrix A with the larger real part
 ρ_f fluid mixture density, kg m⁻³
 σ_i constant coefficient in eq. (8)
 τ $\bar{t}US_v/C_p\rho_f$, dimensionless time
 ψ B/β , Semenov parameter
 $\phi(T)$ $\exp[(T - 1)/\epsilon T]$, temperature dependence of reaction rate constant
 ψ_c critical Semenov number
 Ψ functional defined by (43)

Subscripts and superscripts

- ⁰ initial condition
- ^{∞} limit at unbounded time
- ^{*} quantity evaluated at the maximum temperature

REFERENCES

- Adler, J. and Enig, J. W., 1964, The critical conditions in thermal explosion theory with reactant consumption. *Combust. Flame* 8, 97-103.
- Bilous, O. and Amundson, N. R., 1956, Chemical reactor stability and sensitivity II. Effect of parameters on sensitivity of empty tubular reactors. *A.I.Ch.E. J.* 2, 117-126.
- Boddington, T., Gray, P., Kordylewski, W. and Scott, S. K., 1983, Thermal explosions with extensive reactant consumption: a new criterion for criticality. *Proc. R. Soc. A* 390, 13-30.
- Gray, B. F. and Sherrington, M. E., 1972a, Explosive systems with reactant consumption I. Critical conditions. *Combust. Flame* 19, 435-444.
- Gray, B. F. and Sherrington, M. E., 1972b, Explosive systems with reactant consumption II. Stability. *Combust. Flame* 19, 445-448.
- Hirsch, M. W. and Smale, S., 1974, *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, New York.
- Morbidelli, M. and Varma, A., 1988, A generalized criterion for parametric sensitivity: Application to thermal explosion theory. *Chem. Engng Sci.* 43, 91-102.
- Morbidelli, M. and Varma, A., 1989, A generalized criterion for parametric sensitivity: Application to a pseudohomogeneous tubular reactor with consecutive or parallel reactions. *Chem. Engng Sci.* 44, 1675-1696.
- Rabitz, H. and Smooke, M. D., 1988, Scaling relations and self-similarity conditions in strongly coupled dynamical systems. *J. Phys. Chem.* 92, 1110-1119.
- Thomas, P. H. and Bowes, P. C., 1961, Some aspects of the self-heating and ignition of solid cellulosic materials. *Br. J. Appl. Phys.* 12, 222-229.

Vajda, S., Yetter, R. A. and Rabitz, H., 1990, Effects of thermal coupling and diffusion on the mechanism of H_2 oxidation in steady premixed laminar flames. *Combust. Flame*, **82**, 270-297.

Table 1. Values of the critical Semenov number ψ_c at $\epsilon = 0.1$

B	T^*	Δ	$Re(\lambda_{max})$	Predicted values of ψ_c			
				(a)	(b)	(c)	(d)
7	1.18	-2.54	-0.165	1.020	1.300	192.000	10.500
10	1.24	-2.99	0.014	0.933	1.030	15.000	1.480
20	1.38	-3.78	0.416	0.709	0.731	0.661	0.721
30	1.49	-4.08	0.670	0.611	0.614	0.618	0.607
40	1.47	-1.54	0.830	0.560	0.562	0.562	0.560
50	1.57	-2.31	0.915	0.533	0.533	0.533	0.533

(a) This work.

(b) Lowest estimate by Morbidelli and Varma (1988)

(c) Highest estimate by Morbidelli and Varma (1988)

(d) Estimate of Adler and Enig (1964)

Table 2. Values of the critical Semenov number ψ_c at $\epsilon = 0$

B	θ^*	Δ	$Re(\lambda_{max})$	Predicted values of ψ_c			
				(a)	(b)	(c)	(d)
5	1.46	-3.40	-0.187	0.970	1.130	2.580	2.380
7	2.12	-5.23	0.060	0.907	1.010	1.220	1.090
10	3.01	-9.78	0.560	0.756	0.779	0.794	0.758
20	3.63	-5.12	1.500	0.545	0.545	0.545	0.545
30	2.53	3.85	2.331	0.490	0.490	0.490	0.490

(a) This work.

(b) Lowest estimate by Morbidelli and Varma (1988)

(c) Highest estimate by Morbidelli and Varma (1988)

(d) Estimate of Adler and Enig (1964)

CAPTIONS FOR FIGURES

- Figure 1. Geometric characterization of the local behavior of the explosion system in terms of the trace and the determinant of the Jacobian matrix A . The four regions I, II, III, and IV correspond to stable nodes, stable spirals, unstable spirals, and unstable nodes, respectively. The time step between two consecutive points of the trajectories shown is $\Delta t = 0.2$. The parameter values are $B = 50$, $\epsilon = 0.1$, $\psi = 0.532$ (\sqcup , subcritical behavior), $\psi = 0.533$ (\diamond , critical point), and $\psi = 0.5332$ ($+$, slightly supercritical behavior).
- Figure 2. The larger real part $Re(\lambda_{max})$ of the two eigenvalues λ_1 and λ_2 of the Jacobian matrix A at the temperature maximum T^* as function of the Semenov number ψ at $\epsilon = 0.1$ and three values of B .
- Figure 3. $Re(\lambda_{max})$ at the critical Semenov number ψ_c as function of B . While ψ_c is defined for any B and ϵ as the value of ψ at which $Re(\lambda_{max})$ attains its maximum, it does not imply criticality if $Re(\lambda_{max}) < 0$. Thus, for any ϵ there exists no critical Semenov number below a certain value of B .
- Figure 4. Semi-logarithmic sensitivity functions $\partial T / \partial \log p_j$ of the temperature T at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.55$ (subcritical behavior).
- Figure 5. Semi-logarithmic sensitivity functions $\partial z / \partial \log p_j$ of the conversion z at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.55$ (subcritical behavior).
- Figure 6. Semi-logarithmic sensitivity functions $\partial T / \partial \log p_j$ of the temperature T at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.6107$ (critical point).
- Figure 7. Semi-logarithmic sensitivity functions $\partial z / \partial \log p_j$ of the conversion z at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.6107$ (critical point).
- Figure 8. Semi-logarithmic sensitivity functions $\partial T / \partial \log p_j$ of the temperature T at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.63$ (supercritical behavior).

- Figure 9. Semi-logarithmic sensitivity functions $\partial z / \partial \log p_j$ of the conversion z at $B = 30$, $\epsilon = 0.1$, and $\psi = 0.63$ (supercritical behavior).
- Figure 10. Residual sum of squares defined by (23) - (25), measuring the similarity of the conversion (Q_1) and temperature (Q_2) sensitivity functions at $\epsilon = 0.1$ and $B = 30$.
- Figure 11. Residual sum of squares defined by (23) - (25), measuring the similarity of the conversion (Q_1) and temperature (Q_2) sensitivity functions at $\epsilon = 0.1$ and $B = 50$.
- Figure 12. Constrained temperature (semi-logarithmic) sensitivity functions of the conversion at $\epsilon = 0.1$, $B = 30$, and $\psi = 0.6107$.
- Figure 13. Constrained conversion (semi-logarithmic) sensitivity functions of the temperature at $\epsilon = 0.1$, $B = 30$, and $\psi = 0.6107$.
- Figure 14. Terms in the sensitivity equation for the parameter ψ at $\epsilon = 0.1$, $B = 30$, and $\psi = 0.6107$. Curve 1: $(\partial f_1 / \partial z)(\partial z / \partial \psi) + (\partial f_1 / \partial T)(\partial T / \partial \psi)$. Curve 2: $(\partial f_2 / \partial z)(\partial z / \partial \psi) + (\partial f_2 / \partial T)(\partial T / \partial \psi)$. Curve 3: $\partial f_1 / \partial \psi$. Curve 4: $\partial f_2 / \partial \psi$.

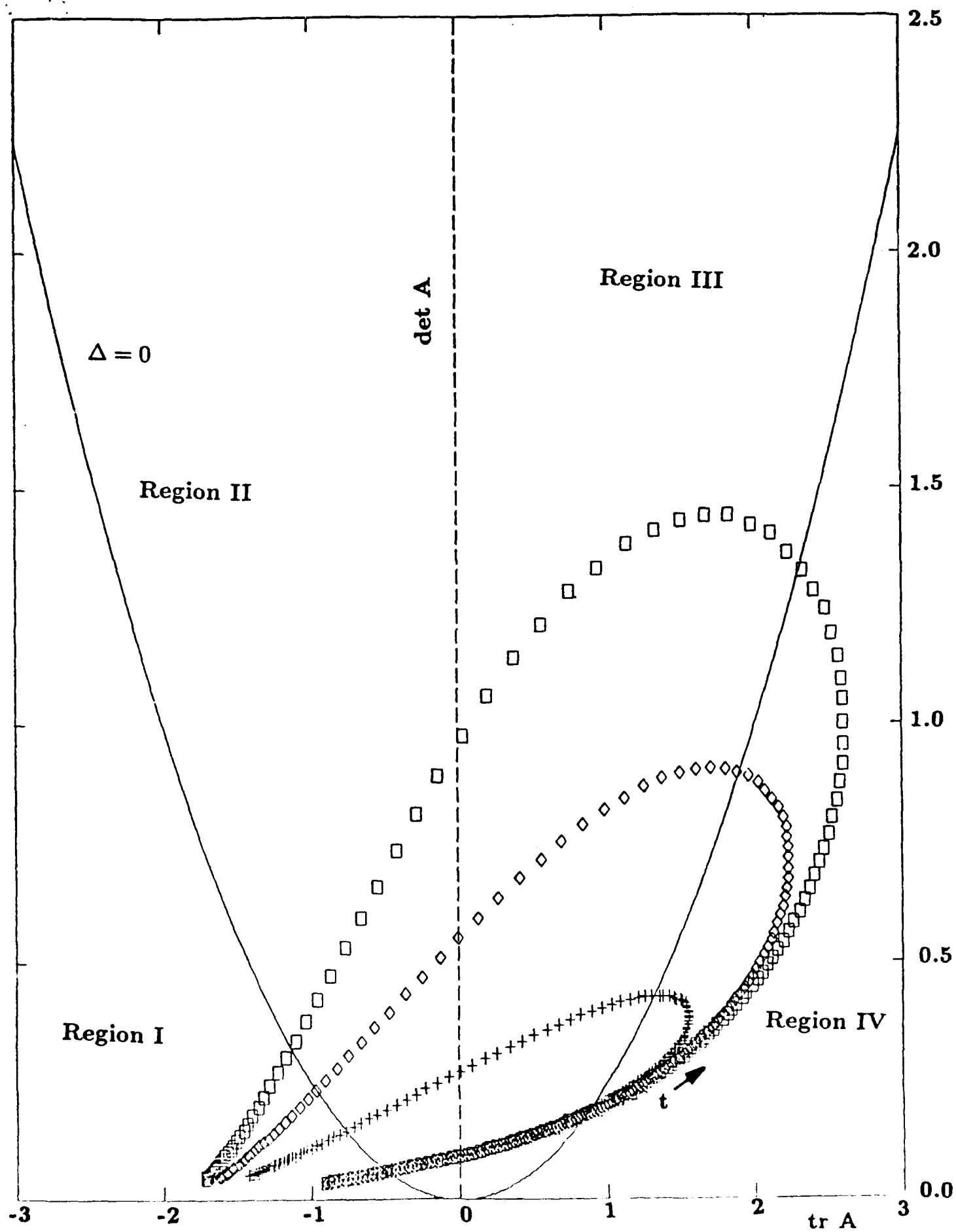


Figure 1

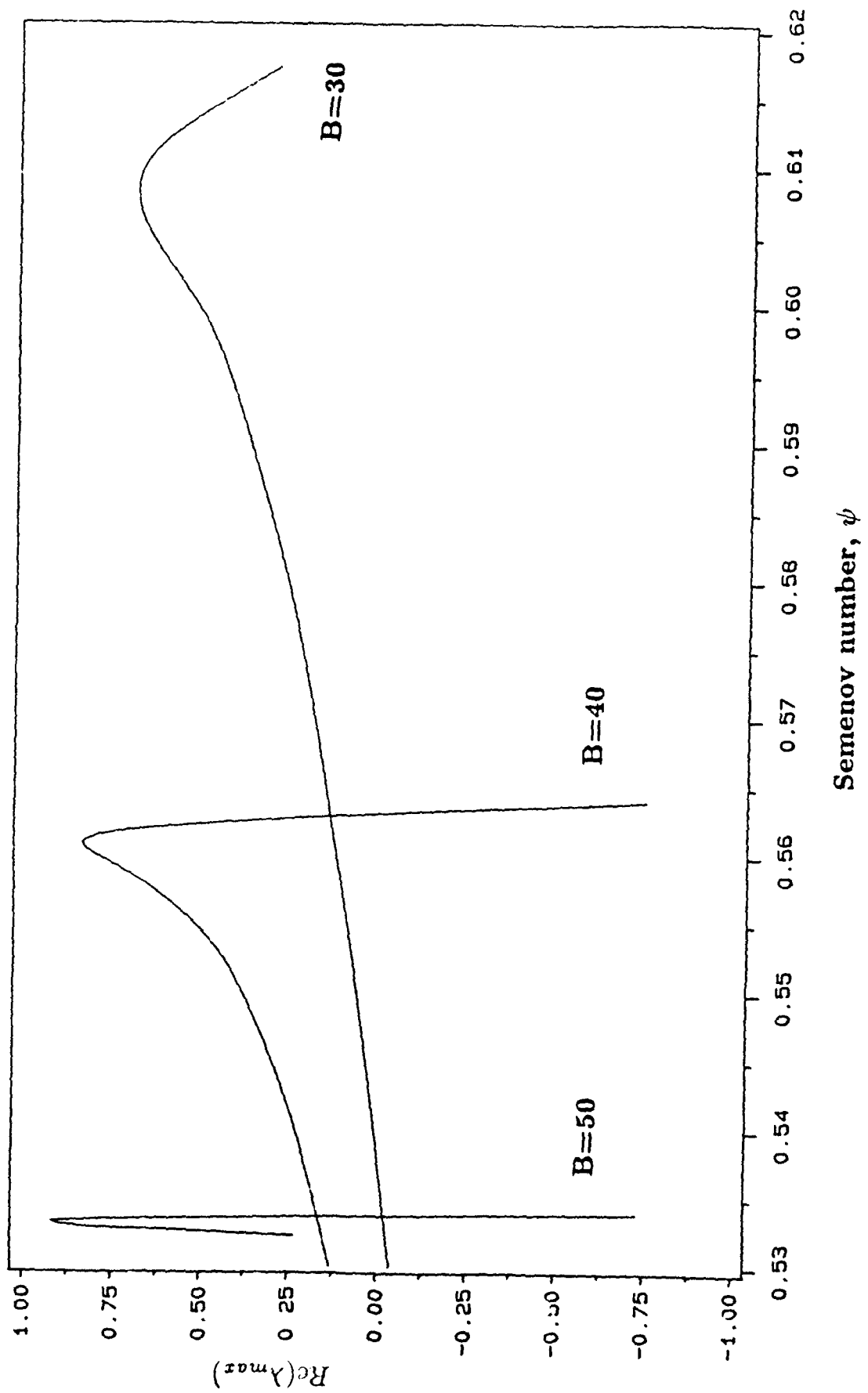


Figure 2

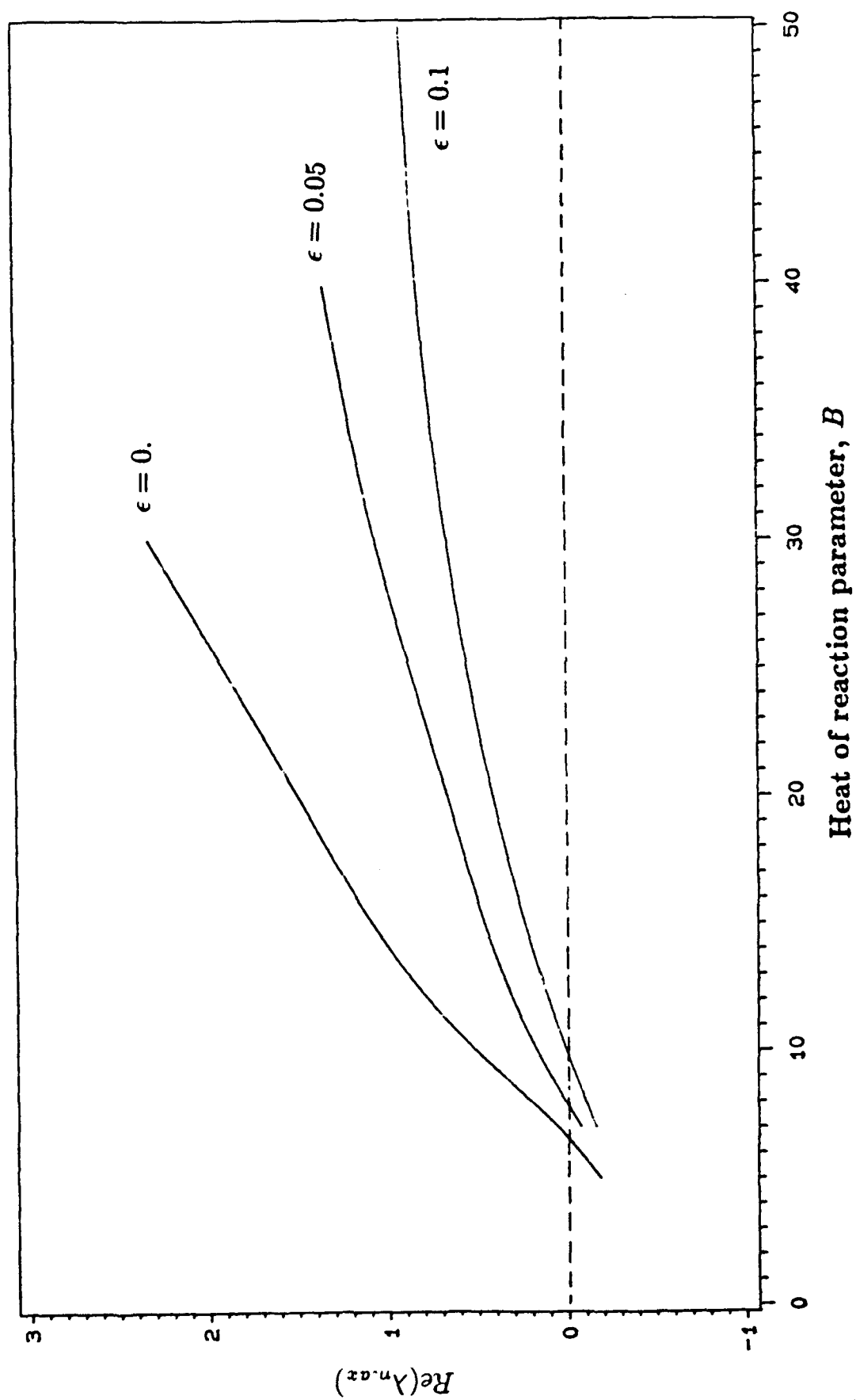


Figure 3

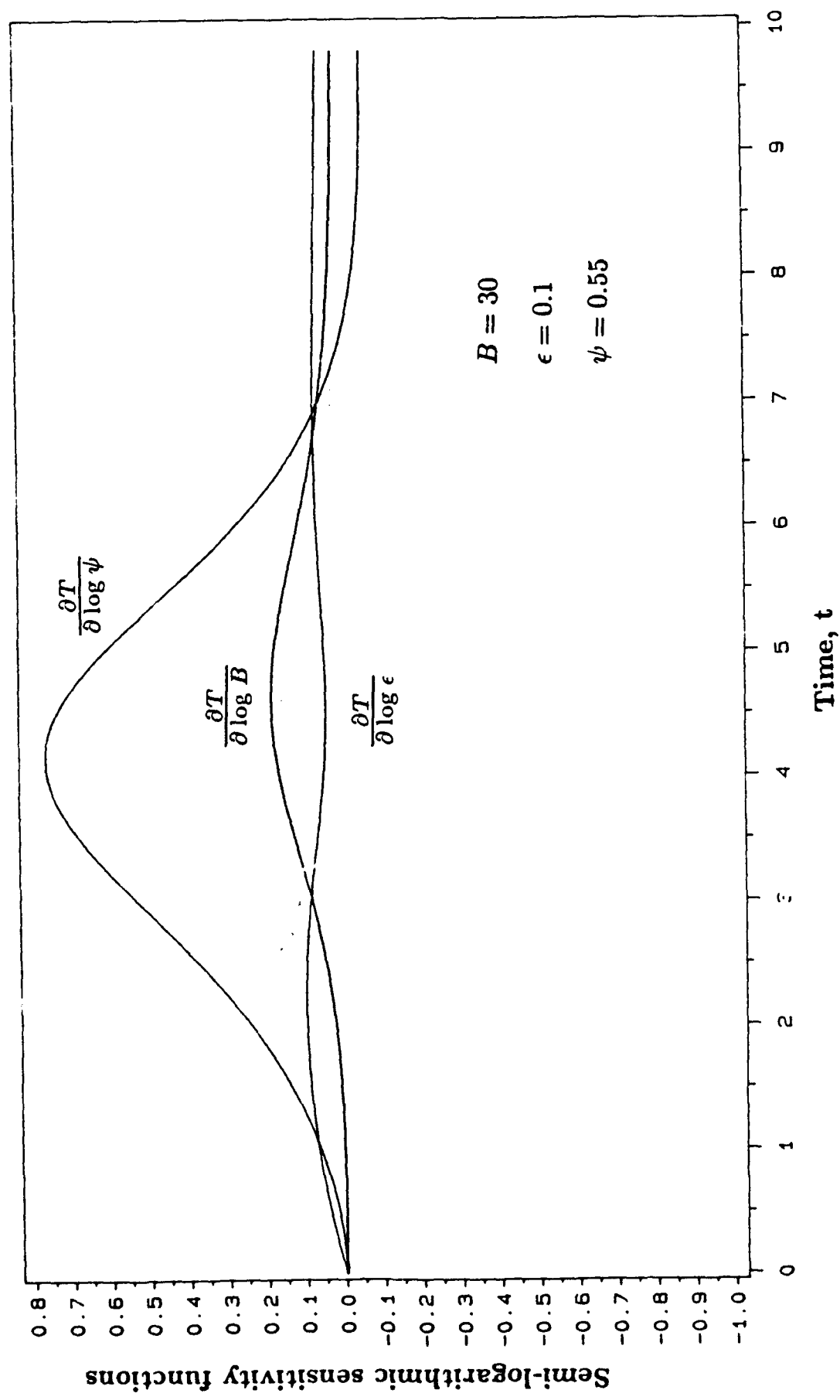


Figure 4

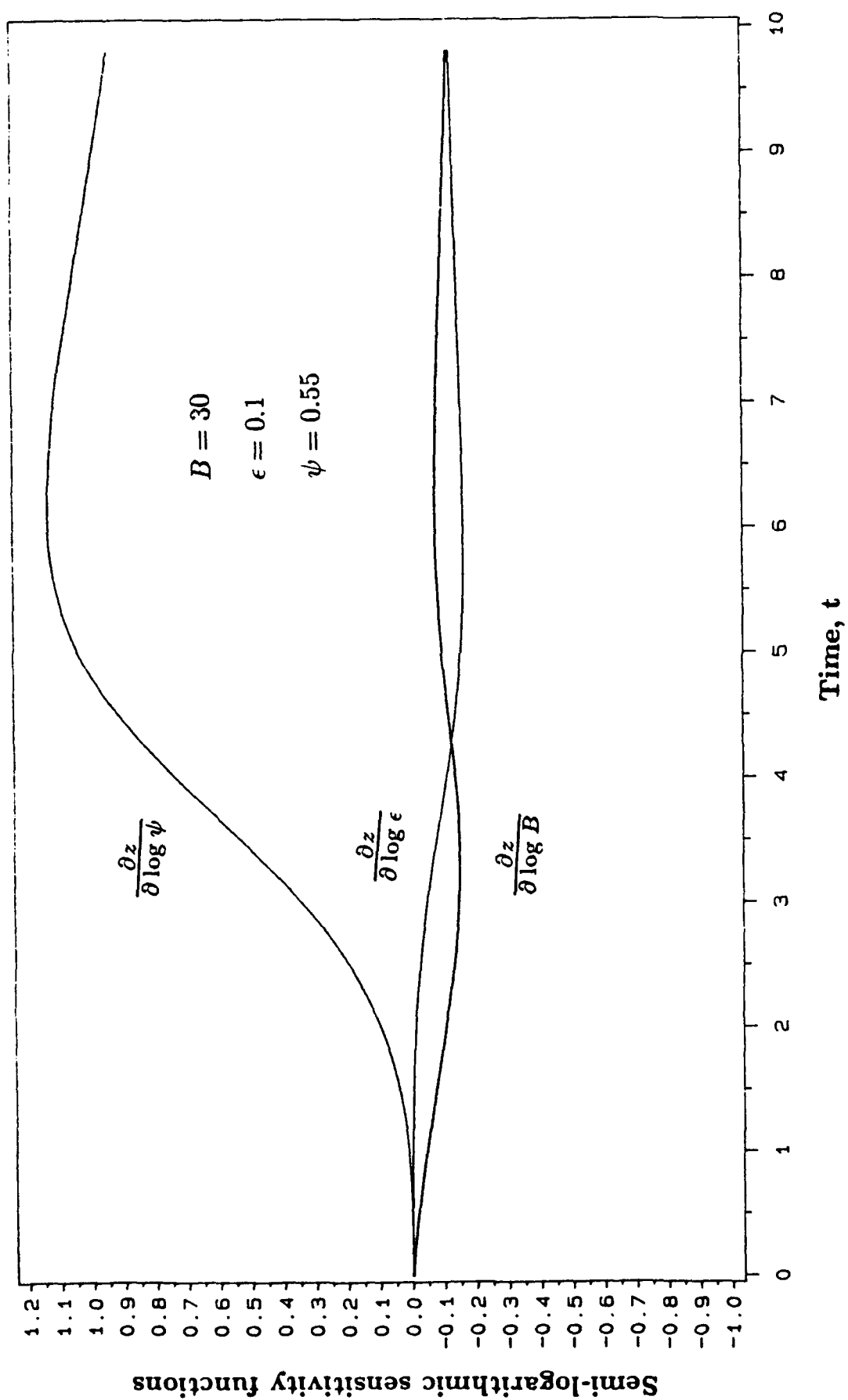


Figure 5

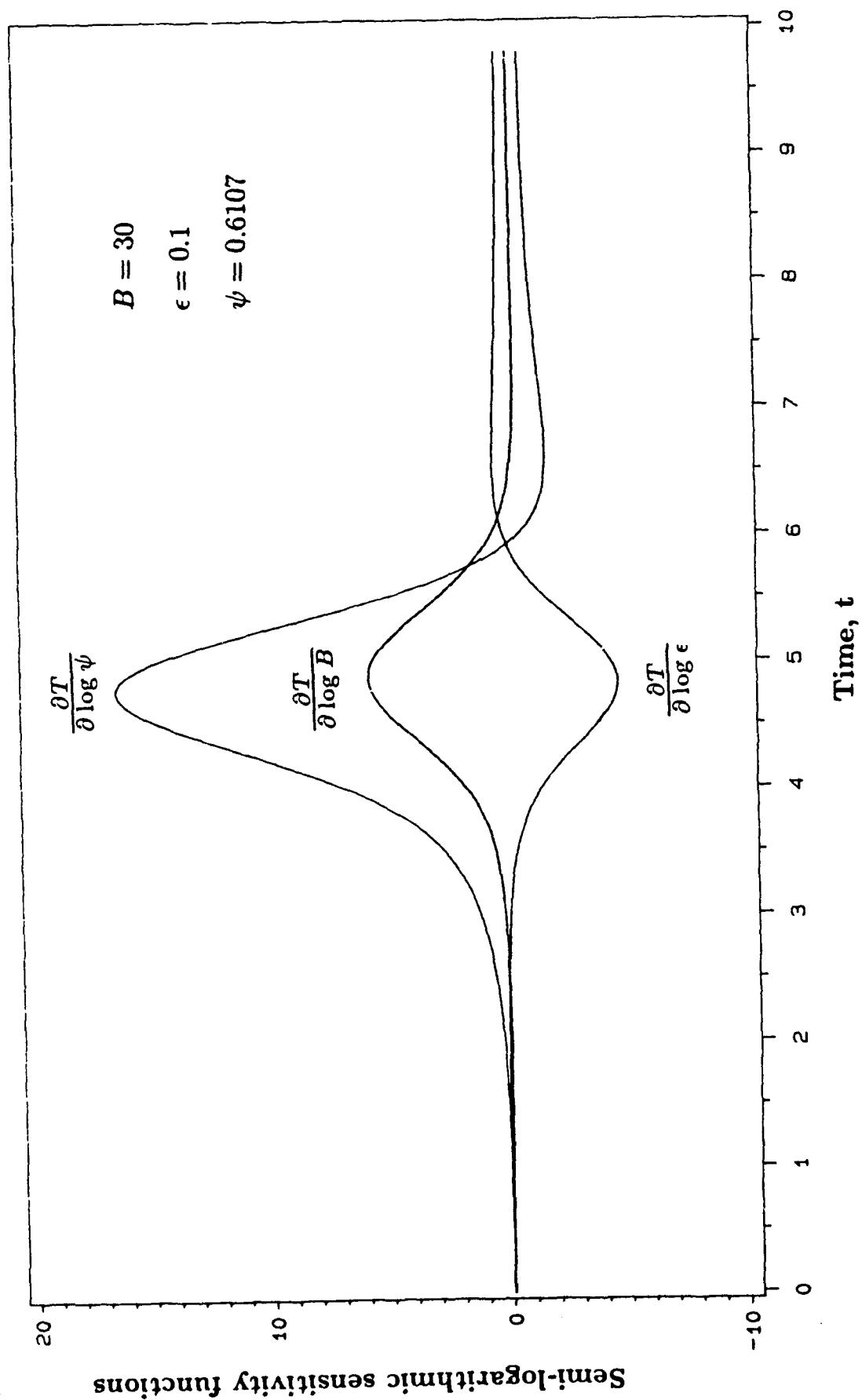


Figure 6

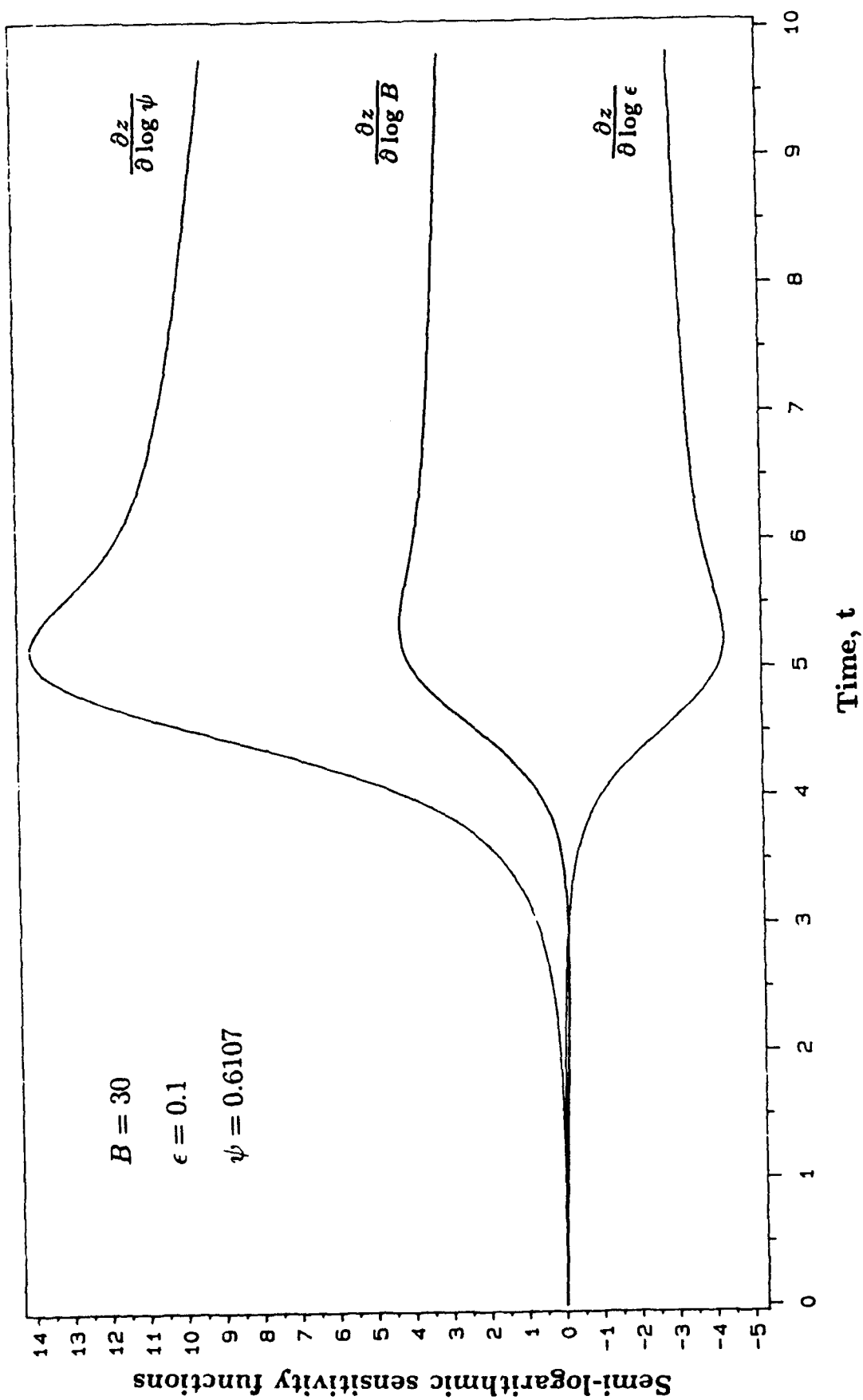


Figure 7

Semi-logarithmic sensitivity functions

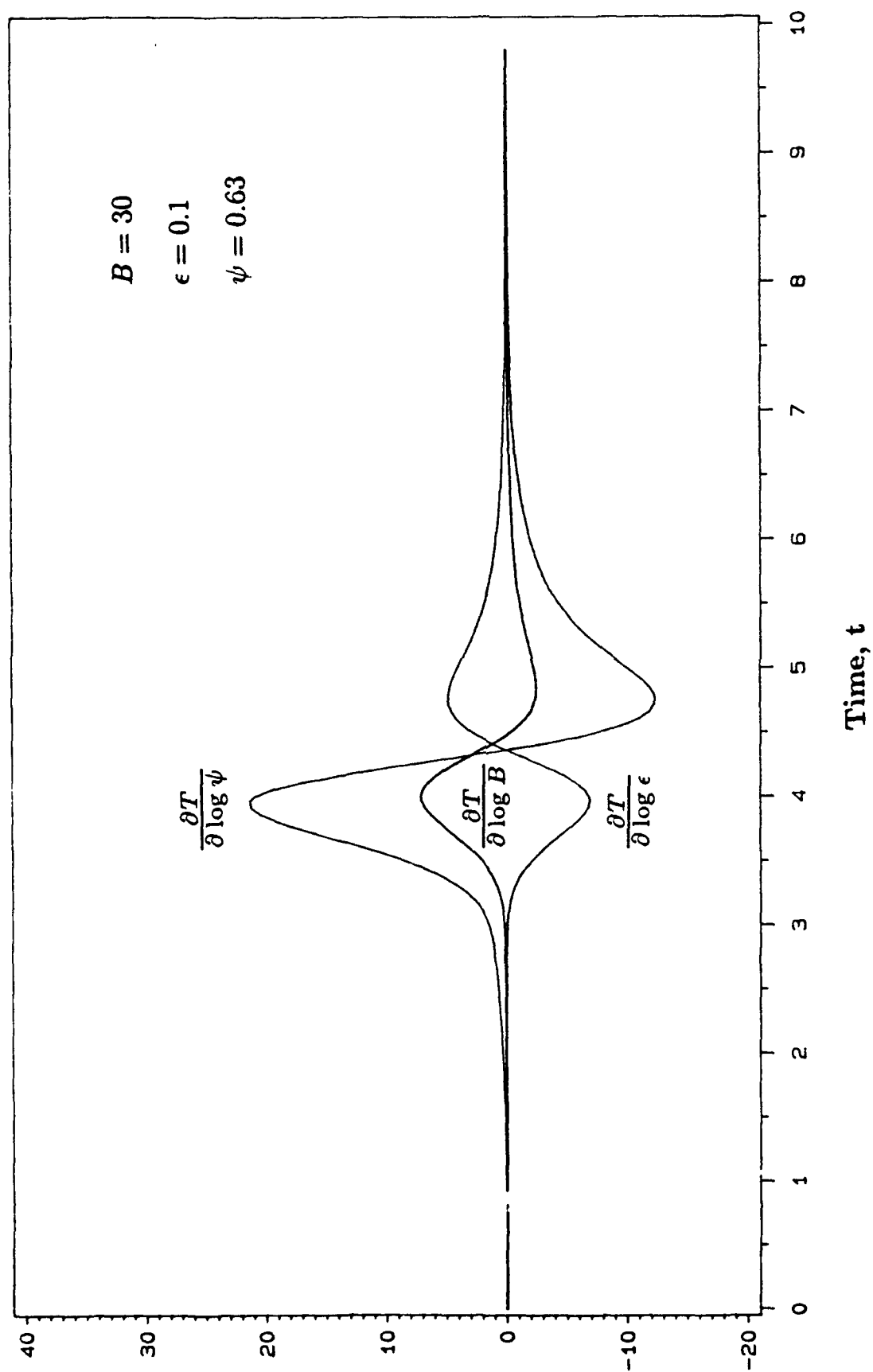


Figure 8

Figure 9

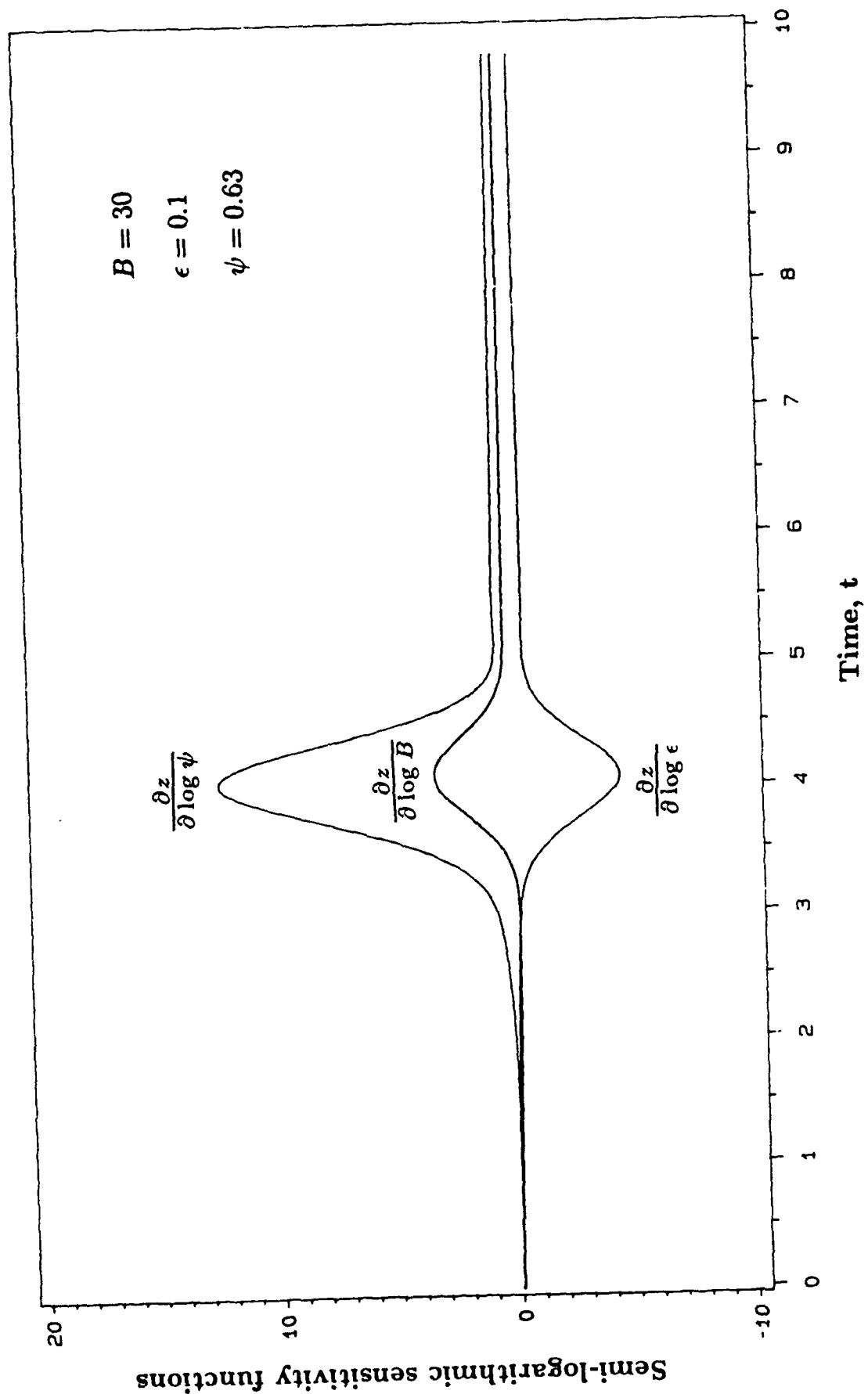
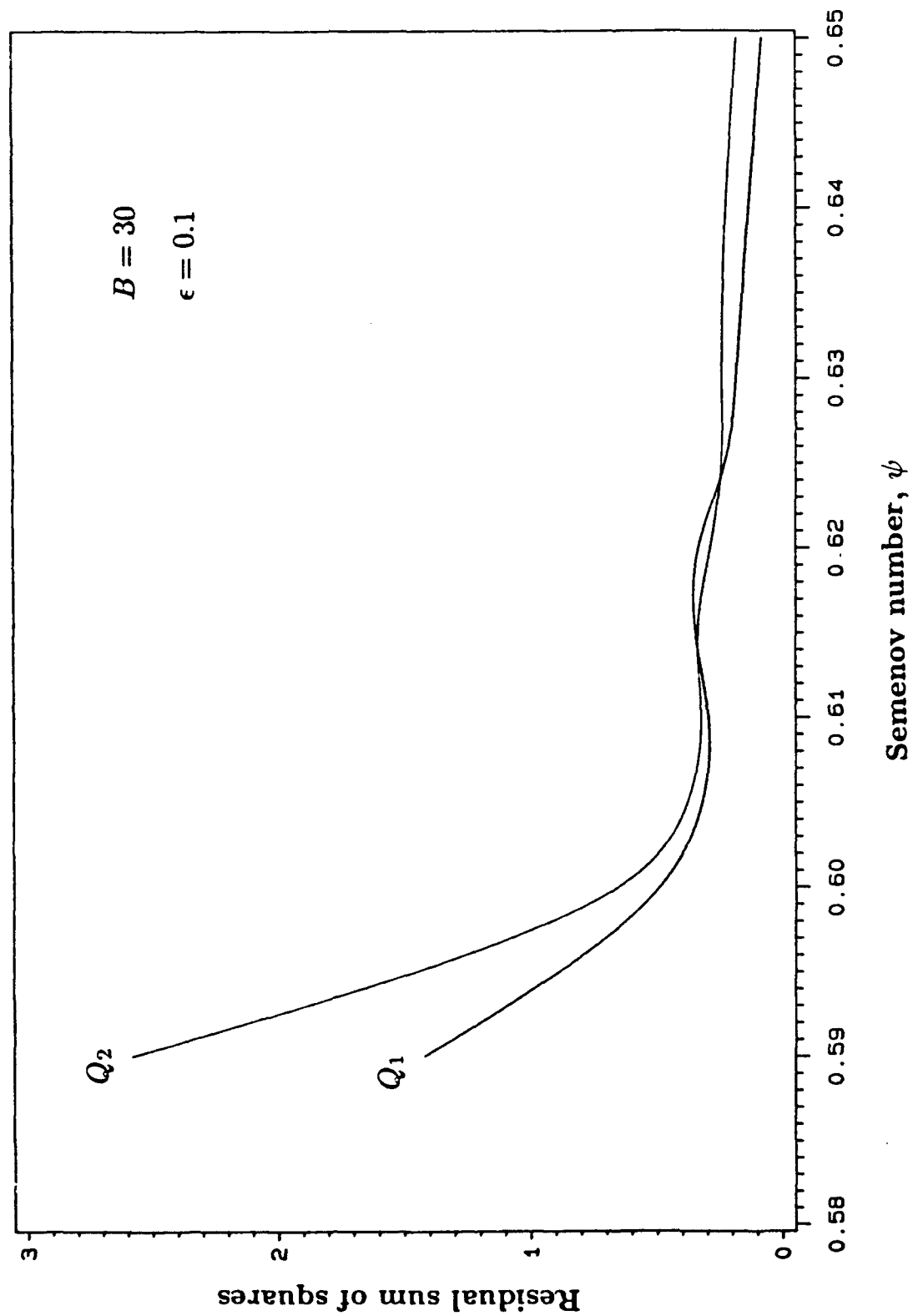


Figure 10



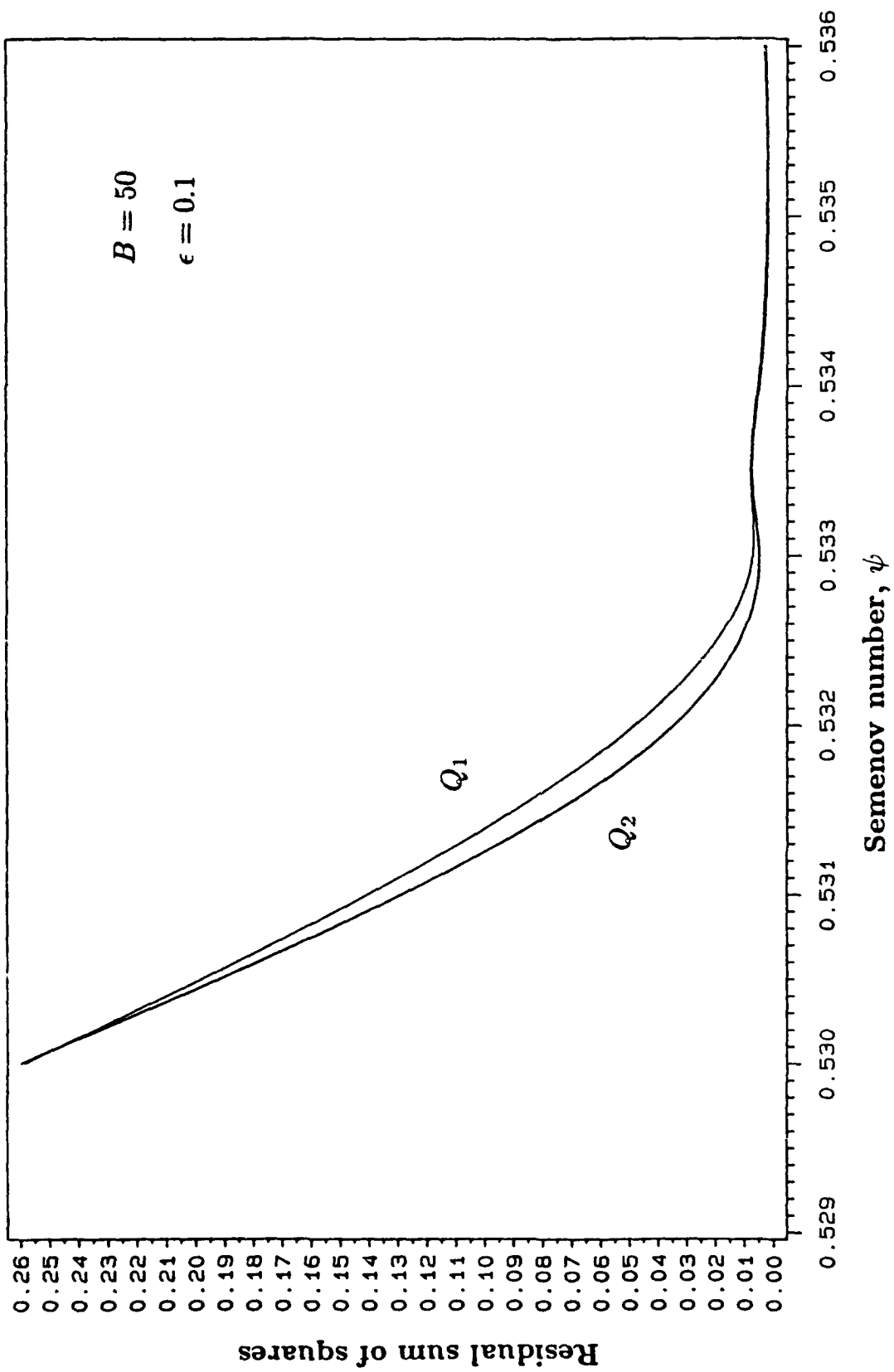


Figure 11

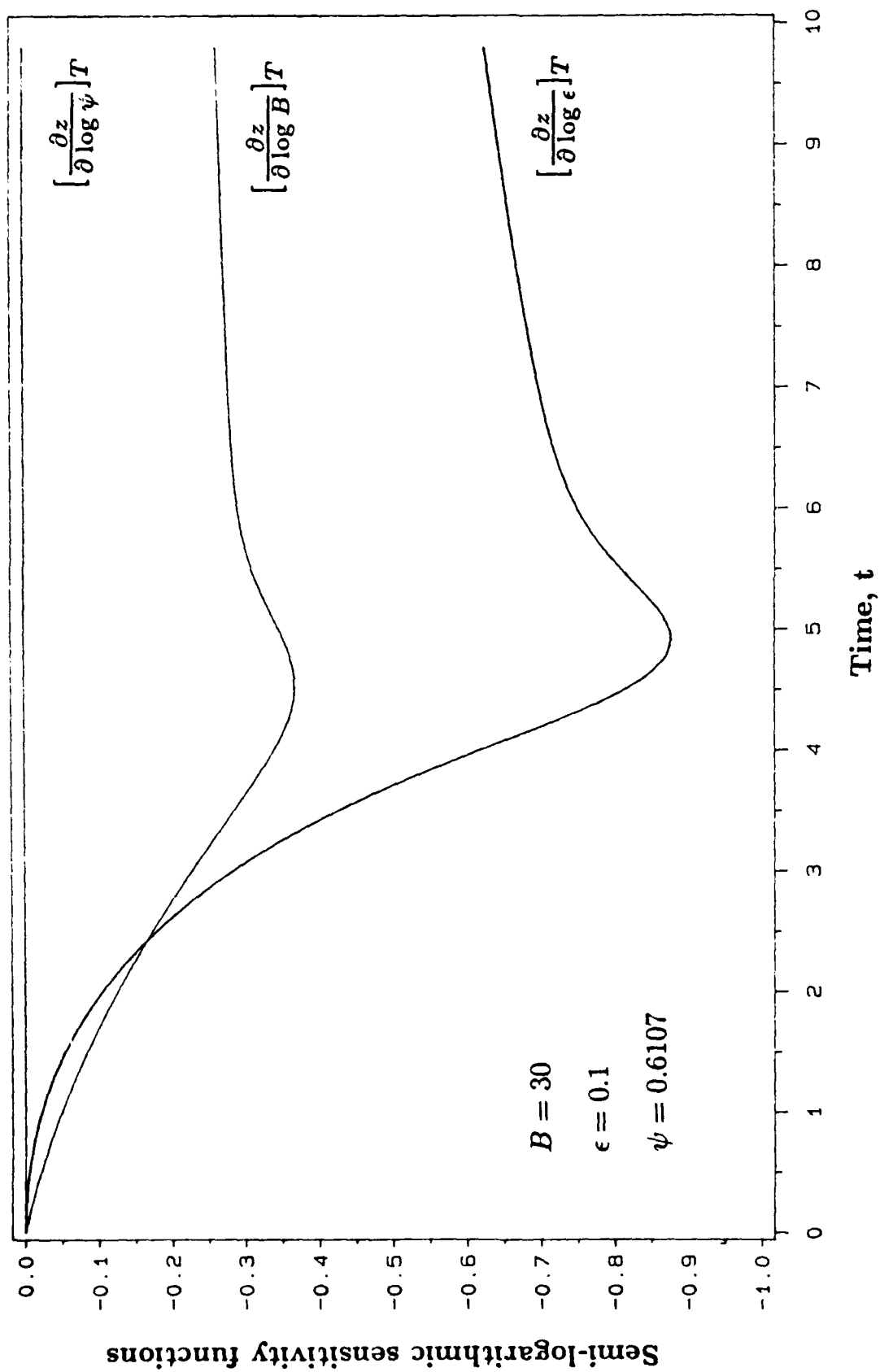


Figure 12

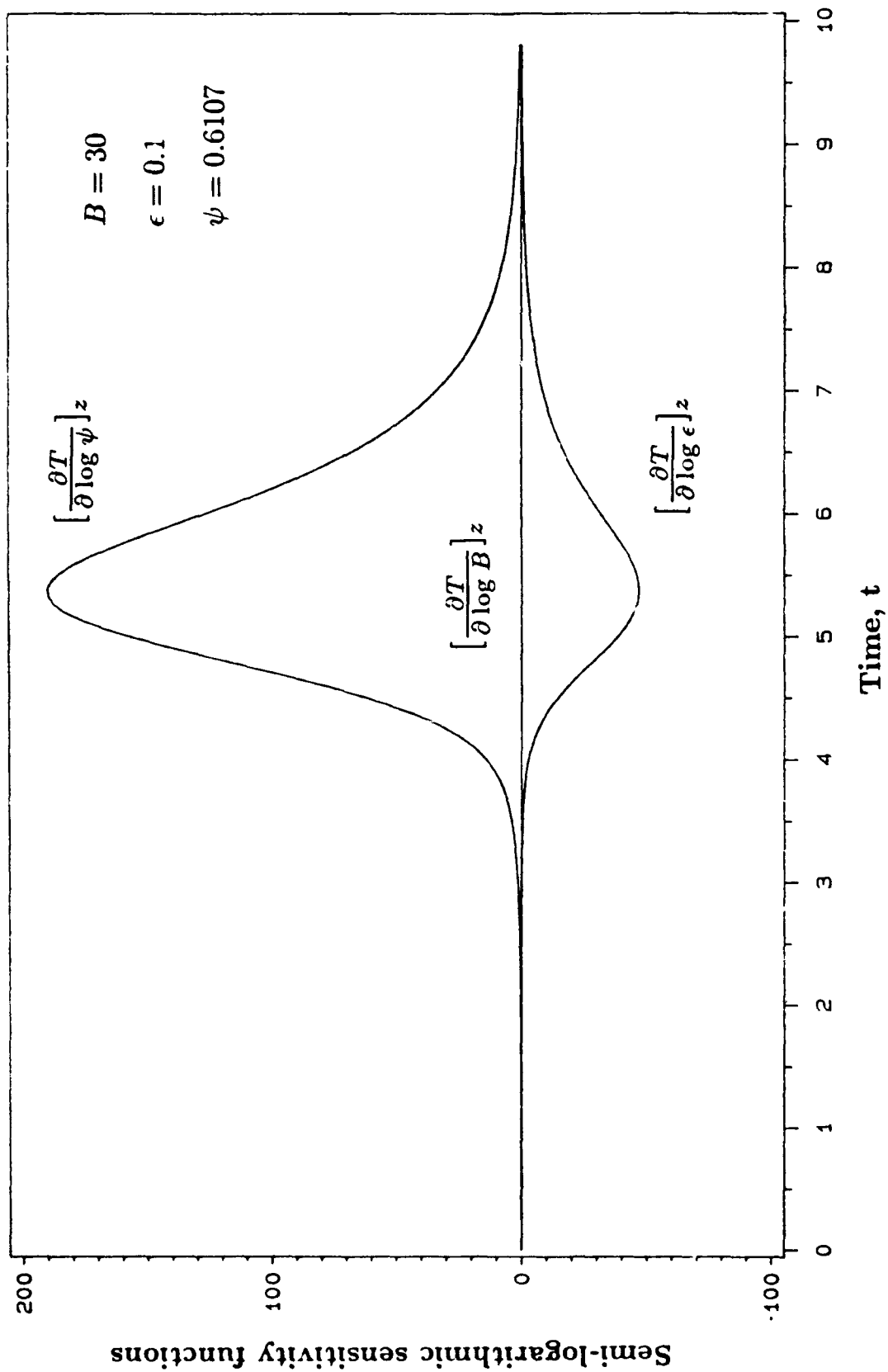


Figure 13

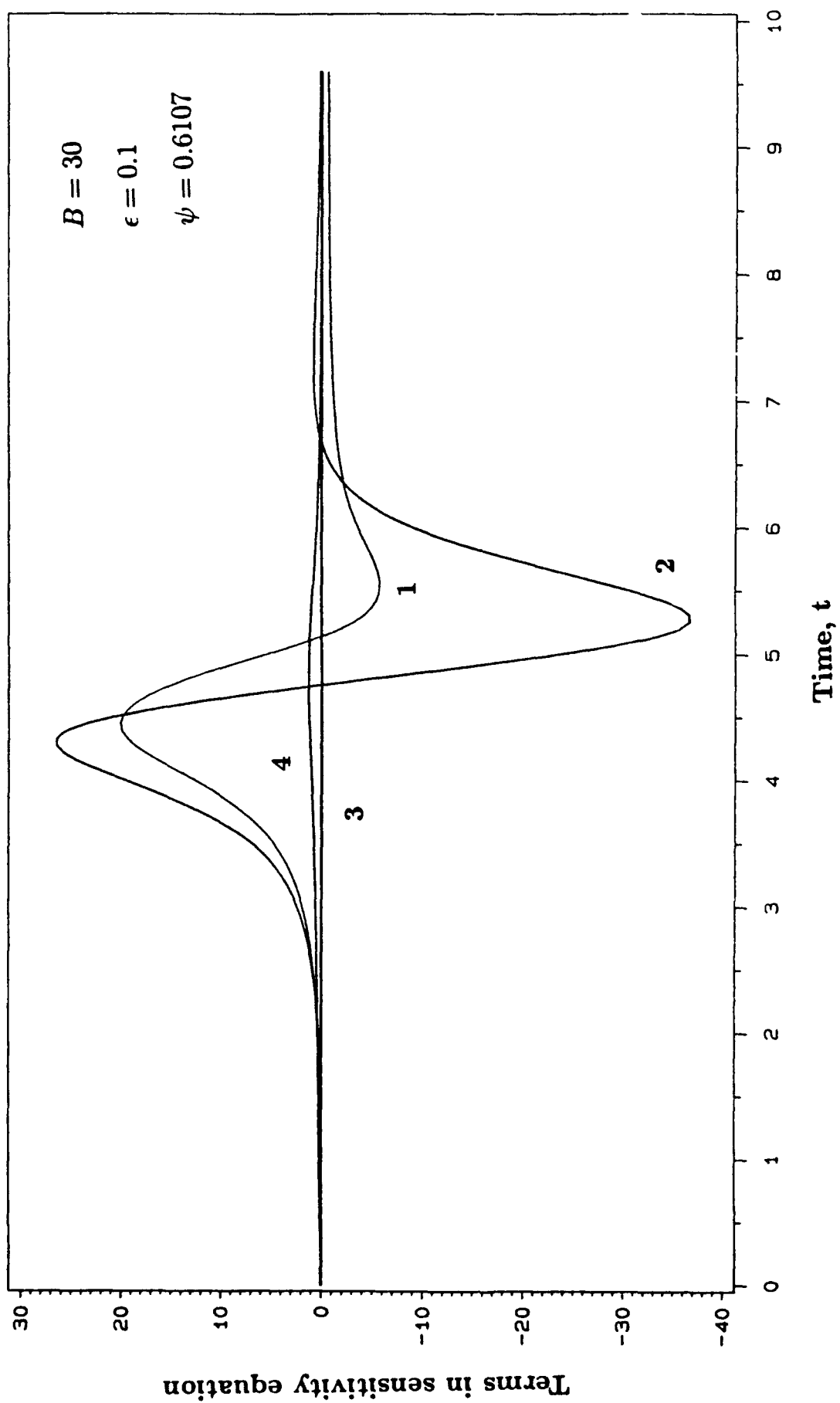


Figure 14

Appendix C

3. A Combined Stability-Sensitivity Analysis of Weak and Strong Reactions of Hydrogen/Oxygen Mixtures, R. Yetter, H. Rabitz and R. Hedges, Int. J. Chem. Kinetics, 23, 51 (1991).

A Combined Stability-Sensitivity Analysis of Weak and Strong Reactions of Hydrogen/Oxygen Mixtures

R. A. YETTER

Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey 08540

H. RABITZ and R. M. HEDGES

Department of Chemistry, Princeton University, Princeton, New Jersey 08540

Abstract

Stability and sensitivity analysis are used to examine the ignition/reaction characteristics of dilute hydrogen-oxygen mixtures. The analysis confirms the existence of two distinct regions of ignition and fast reaction previously labeled "weak" and "strong" ignition, both of which are located in the explosive pressure-temperature domain and separated by a region related to the "extended" classical second limit. The stability analysis is based on an eigenanalysis of the Green's function matrix of the governing kinetic equations. The magnitudes of the largest (and system controlling) eigenvalue allow the strengths of the two processes to be quantified, giving a clear definition to the terms "weak" and "strong." The sensitivities of the largest eigenvalue to the reaction rate constants of the mechanism pinpoint the elementary steps controlling the two ignition processes and the subsequent reaction. The associated eigenvectors yield the direction of change in species concentrations and temperature during the course of reaction. These vectors are found to be nearly constant during the induction period of both "weak" and "strong" ignition, thus producing constant overall stoichiometric reactions. The subsequent reaction of major reactants associated with "weak" ignition also has a constant overall reaction vector, although, different than that during the induction period. However, the vector describing the reaction of major reactants associated with "strong" ignition is found never to be constant, but continuously changing beyond the induction period.

Introduction

Advanced flight concepts such as the aerospace plane have renewed interest in air breathing hypersonic combustion. Hydrogen, because of its high specific energy and high capacity for cooling, is a prime candidate to fuel these propulsion systems. Because of short residence times in such combustors, fundamental understanding of hydrogen-oxygen ignition and stability characteristics are essential for proper combustor design and practical implementation of hydrogen as a fuel.

Hydrogen-oxygen kinetics have been observed to exhibit significantly different ignition characteristics depending upon the initial pressure and temperature of an explosive mixture. The differences in behavior, termed "strong" (or "sharp") ignition and "weak" (or "mild") ignition, were first noted by Soloukhin and Strehlow [1] and subsequently studied by others [2-5]. Voevodsky and Soloukhin [5] explained these differences by a change in

chemical mechanism which occurs as a result of the "extended" second limit of the classical pressure-temperature explosion limits of H_2/O_2 mixtures. Although this work could not accurately predict the experimental trends of shock induced reactions, the qualitative trends produced from model analysis (based on an inadequate chemical mechanism) were consistent with experiment and implied that the reaction changed from a fully-branched mechanism ("strong" ignition) to a straight-chain mechanism with rare branchings ("weak" ignition).

Meyer and Oppenheim [6], using reflected shock wave data, in addition to Voevodsky and Soloukhin's data, have shown that the separation between weak and strong ignition, although affected by the change in chemistry across the "extended" second limit does not correspond to it, but to a curve represented by the sensitivity of the induction time to the initial temperature $\partial\tau/\partial T = -2 \mu s/K$. They argued that weak ignition delays are very sensitive to gas dynamic disturbances, such as perturbations in the temperature field, whereas strong ignition delays were not, thus altering the weak-strong ignition limit from the "extended" second limit.

More recently, Oran and Boris [7] have examined weak and strong ignition numerically with a detailed chemical reaction mechanism more representative of current understanding of H_2/O_2 kinetics than previous analysis. Their results were consistent with the ideas of Meyer and Oppenheim and also showed that the ignition process is strongly sensitive to sound wave and entropy (temperature) perturbations. Their work did not, however, conclusively determine the criteria of weak and strong ignition.

The present paper reexamines the H_2/O_2 ignition process using stability and sensitivity analysis techniques, which are shown to yield further understanding of the chemistry of this process. The details of the analysis procedure have been described previously [8]. However, this article extends the methodology to include the case of degeneracy among eigenvalues.

Reaction Model

The reaction mechanism used in this analysis, given in Tables I and II, includes 9 chemical species and 19 forward and reverse elementary reactions and is based on a reaction mechanism originally developed and validated for $CO/H_2/O_2$ kinetics [9]. All of the thermochemical data are from the JANAF tables [10] with the exception of the heat of formation for HO_2 , which is from Shum and Benson [11]. The temperature dependencies of the thermochemical data are stored as polynomial fits in the format of the NASA chemical equilibrium program [12]. The polynomial coefficients for all species, except HO_2 , are from Kee et al. [13]. The polynomial coefficients for HO_2 were obtained using the THERM code [14]. Rate constants, obtained from literature evaluations, are specified for one direction only. Thermochemical data are used to evaluate the reverse reaction rate constants. Chaperon efficiencies are used for the dissociation/recombination reactions as specified in Table II. This mechanism differs from that originally developed for $CO/H_2/O_2$ kinetics in the heat of formation of HO_2 (in ref. [9], $\Delta H_{f,298}^\circ = 3.0 \pm 0.4$ kcal/mol [15]) and in the rate constant expressions for reactions 14 and 15. The rate constant for the $HO_2 + HO_2 \rightarrow H_2O_2 + O_2$ reaction [16] is expressed as a double exponential to account for

TABLE I. $\Delta H_f(298)$, $S(298)$, and $C_p(T)$ for atomic and molecular species considered in the H_2/O_2 reaction (units are kcal-mol $^{-1}$ -K).

Species	$\Delta H_f(298)$	$S(298)$	$C_p(300)$	$C_p(500)$	$C_p(800)$	$C_p(1000)$	$C_p(1500)$	$C_p(2000)$
H	52.103 \pm 0.001	27.416 \pm 0.004	4.97	4.97	4.97	4.97	4.97	4.97
O	59.55 \pm 0.024	38.49 \pm 0.005	5.23	5.08	5.02	5.00	4.98	4.96
OH	9.318 \pm 0.29	43.905 \pm 0.01	7.15	7.07	7.13	7.33	7.87	8.28
H ₂	0	31.232 \pm 0.008	6.90	7.00	7.07	7.21	7.73	8.18
O ₂	0	49.03 \pm 0.008	7.01	7.44	8.07	8.35	8.72	9.03
H ₂ O	-57.795 \pm 0.01	45.13 \pm 0.01	8.00	8.45	9.22	9.87	11.26	12.22
HO ₂	3.5 \pm 0.5	54.42 \pm 0.02	8.36	9.48	10.75	11.37	12.34	12.90
H ₂ O ₂	-32.53	55.66	10.42	12.35	14.29	15.21	16.85	17.88
N ₂	0	45.93 \pm 0.005	6.95	7.08	7.50	7.83	8.32	8.60

TABLE II. H₂/O₂ reaction mechanism (reaction rates in cm³-mol-s-kcal units, $k = AT^n \exp(-E_a/RT)$ unless specified).

	ΔH_{298}	$\log(A_f)$	n_f	$E_{a,f}$	UF	T_{RANGE}	Reference
H₂-O₂ Chain Reactions							
1. H + O ₂ = O + OH	16.77	14.28	0.00	16.44	2	962-2577 K	Pirraglia, et al., (1989) [18]
2. O + H ₂ = H + OH	1.87	4.71	2.67	6.29	1.5	297-2495 K	Sutherland, et al., (1986) [19]
3. OH + H ₂ = H + H ₂ O	-15.01	8.33	1.51	3.43	1.5	250-2581 K	Michael & Sutherland, (1988) [20]
4. OH + OH = O + H ₂ O	-16.88	$k = 5.46 \times 10^{11} \exp(0.00149 \times T)$			2.5	250-2000 K	Tsang & Hampson (1986) [21]
H₂-O₂ Dissociation/Recombination Reactions							
5. H ₂ + M = H + H + M (N ₂) ^a	104.2	19.66	-1.40	104.38	3	600-2000 K	Tsang & Hampson (1986) [21]
6. O + O + M = O ₂ + M (N ₂)	-119.1	15.79	-0.50	0.00	1.3	2000-10000 K	Tsang & Hampson (1986) [21]
7. O + H + M = OH + M	-102.3	18.67	-1.00	0.00	10	—	Tsang & Hampson (1986) [21]
8. H + OH + M = H ₂ O + M(N ₂)	-119.2	22.35	-2.00	0.00	2	1000-3000 K	Tsang & Hampson (1986) [21]
Formation and Consumption of HO₂							
9. H + O ₂ + M = HO ₂ + M(N ₂)	-48.60	19.83	-1.42	0.00	3	200-2000 K	Slack (1977) [22]
10. HO ₂ + H = H ₂ + O ₂	-55.60	13.82	0.00	2.13	2	298-773 K	Tsang & Hampson (1986) [21]
11. HO ₂ + H = OH + OH	-36.97	14.23	0.00	0.87	2	298-773 K	Tsang & Hampson (1986) [21]
12. HO ₂ + O = OH + O ₂	-53.73	13.24	0.00	-0.40	1.2	200-400 K	Tsang & Hampson (1986) [21]
13. HO ₂ + OH = H ₂ O + O ₂	-70.61	16.16	-1.00	0.00	2	298-1400 K	Tsang & Hampson (1986) [21]
Formation and Consumption of H₂O₂							
14. HO ₂ + HO ₂ = H ₂ O ₂ + O ₂	-39.53	$k = 1.08 \times 10^{11} \exp(+1759/RT)$ $+ 1.26 \times 10^{14} \exp(-10038/RT)$			3	298-1100 K	Lightfoot, et al. (1988) [16]
15. OH + OH = H ₂ O ₂ (N ₂) ^b	-51.17	$k_0 = [M] \times 2.90 \times 10^{17} (T/300)^{-0.76}$ $k_{\text{inf}} = 9.12 \times 10^{12} (T/300)^{-0.37}$ $F_{\text{cent}} = 0.5, N = 1.13$					
16. H ₂ O ₂ + H = H ₂ O + OH	-68.05	13.00	0.00	3.59	2	700-1500 K	Brouwer, et al. (1987) [17]
17. H ₂ O ₂ + H = H ₂ + HO ₂	-16.07	13.68	0.00	7.95	3	283-800 K	Warnatz (1985) [23]
18. H ₂ O ₂ + O = OH + HO ₂	-14.20	6.98	2.00	3.97	5	283-800 K	Tsang & Hampson (1986) [21]
19. H ₂ O ₂ + OH = H ₂ O + HO ₂	-31.08	12.85	0.00	1.43	3	250-800 K	Tsang & Hampson (1986) [21]
					2	298-800 K	Warnatz (1985) [23]

^a(N₂) [M] = [N₂] + [H] + [O] + [OH] + 2.5[H₂O] + [O₂] + [HO₂] + [H₂O₂].^b $k_{15} = k_{\text{inf}} [k_0/k_{\text{inf}} / (1 + k_0/k_{\text{inf}})] F_{\text{cent}} \{1 + [\log(k_0/k_{\text{inf}})/N]^2\}^{-1}$.UF = uncertainty factor, $k_{\text{min}} = k/\text{UF}$ and $k_{\text{max}} = k \times \text{UF}$.

a negative activation energy observed at low temperatures ($T < 700$ K) due to an association process and for a positive activation energy observed at high temperature ($T > 700$ K) due to an abstraction process. For the pressure-dependent rate constant of $\text{OH} + \text{OH} \leftrightarrow \text{H}_2\text{O}_2$ [17], fall-off behavior has been included and expressed in the Troe formulation.

The equations for a constant volume mixture reacting homogeneously are

$$(1) \quad \frac{dC_i}{dt} = \dot{\omega}_i, \quad C_i(t_0) = C_{i,0} \quad i = 1, \dots, N - 1$$

$$(2) \quad \frac{dT}{dt} = \sum_{i=1}^{N-1} (h_i - RT)\dot{\omega}_i / \sum_{i=1}^{N-1} C_{v,i} C_i, \quad T(t_0) = T_0$$

where C_i is the molar concentration of the i -th chemical species, $\dot{\omega}_i$ is the molar production rate of the i -th chemical species, T is the mixture temperature, $C_{v,i}$ is the specific heat at constant volume of the i -th chemical species, h_i is the enthalpy of the i -th chemical species, and t is time. The kinetic equations are solved numerically using LSODE [24] and CHEMKIN [25].

This system of equations is a good approximation for describing the kinetics of many experiments, including static reactors and shock tubes. The present chemical model does not include surface kinetics nor does the mathematical model have spatial dependence, and hence, the findings reported here are based "purely" on gas-phase kinetics.

A comparison of ignition delays between model prediction and experimental measurement is given in Table III. The experiments are those of Skinner and Ringrose [26] who studied ignition delays of a mixture consisting of 8% H_2 and 2% O_2 in argon which were heated behind reflected shocks to temperatures between 964 and 1075 K and a pressure of 5 atm. For the calculations, the rate constants used for the pressure dependent reactions with Ar as the collision partner are those reported in refs. [9] and [17]. The ignition delay is defined here as the reaction time to the maxima in OH concentration. In Table IV, another set of comparisons between experimental and calculated ignition delays are presented for higher temperatures and a lower pressure. The experiments are those of Schott and Kinsey [27] who studied ignition delays of a mixture consisting of 1% H_2 and 2% O_2 in argon which were heated behind incident shock waves to temperatures between 1082 and 1836 and a pressure of 1 atm. The ignition delay is defined here as the time required for the OH concentration to equal 1×10^{-9} mol/cm³. Overall, the agreement is observed to be better for the set of data at higher temperatures than the data set at lower temperatures. The reported differences in ignition delay data may result from both experimental and model uncertainties. Indeed, accurate measurements of absolute ignition delay times are difficult, as is evident from the reproducibility of the data themselves. Based on an overall activation energy obtained from the low temperature experiments, we note that at 1000 K, an uncertainty of even 25 K in T_0 results in an uncertainty of a factor of approximately 3 in ignition delay. Such an uncertainty in T_0 is likely in the present experiments. Lastly, note that the agreement between model and experiment is generally within the uncertainties of the individual rate constants of the mechanism (see Table II). A discussion on the most sensitive reactions of the mechanism is included below.

TABLE III. Induction times for gas mixture containing 8% H₂ and 2% O₂ in argon at 5 atm total pressure

$T(K)$	$\tau^e(ms)$	$\tau^m(ms)$
964	15.0	25.9
965	10.0	25.0
981	4.3	14.4
1004	1.7	6.6
1005	2.3	6.4
1024	0.9	3.2
1075	0.22	0.36

τ -induction time is defined as the time required to reach the maxima in OH concentration.

e -experimental measurements (from Skinner and Ringrose [26]).

m -model prediction.

TABLE IV. Induction times for gas mixture containing 1% H₂ and 2% O₂ in argon at 1 atm total pressure.

$T(K)$	$\tau^e(\mu s)$	$\tau^m(\mu s)$
1082	570	857
1085	630	838
1154	330	521
1180	340	441
1200	300	394
1275	310	264
1292	174	242
1304	140	229
1305	161	228
1310	185	222
1313	175	219
1615	55	70
1625	66	68
1644	58	64
1666	59	60
1825	40	39
1836	37	38

τ -induction time is defined as the time required for the OH concentration to equal 1×10^{-9} mol/cm³.

e -experimental measurements (from Schott and Kinsey [27]).

m -model prediction.

This mechanism has also been compared with experimental data from flow reactor experiments [9,28], which have tested the kinetics during the consumption of major reactants, assuming constant pressure and adiabatic conditions. The comparisons, made between time dependent H₂ and O₂ concentration profiles and the temperature profile for dilute mixtures reacting in N₂ at 910 K and 1 atmosphere, were found to be good.

Green's Function Stability and Sensitivity Analysis

The constant volume model described above can be rewritten in simplified notation as

$$(3) \quad \frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}), \quad \mathbf{X}(t_0) = \mathbf{X}_0$$

where the dependent vector \underline{X} consists of the species concentrations and the system temperature.

The Green's function of this differential equation system arises from a linearization about a time-dependent reference solution (and not about a point in the solution). It satisfies the matrix differential equation

$$(4) \quad \frac{d\underline{G}(t, t_o, \underline{X})}{dt} = \underline{J}[\underline{X}(t)]\underline{G}(t, t_o, \underline{X}), \underline{G}(t_o, t_o, \underline{X}) = \underline{1}$$

where \underline{J} is the $N \times N$ Jacobian matrix of the system equations with elements $J_{ij} = \partial F_i / \partial X_j$. The \underline{X} dependence of \underline{G} indicates that it is functionally dependent upon the entire reference trajectory over the interval $t_o \rightarrow t$. The formal solution of eq. (4) is

$$(5) \quad \underline{G}(t, t_o, \underline{X}) = \underline{T} \exp \left[\int_{t_o}^t \underline{J}(t') dt' \right]$$

where \underline{T} is a time ordering operator [29].

The Green's function of the solution can be interpreted as the sensitivity of the differential equation system to the initial conditions [30],

$$(6) \quad G_{ij} = \frac{\partial X_i(t)}{\partial X_j(t_o)}$$

The ij -th element of the matrix prescribes how the i -th component of \underline{X} changes at time t when the j -th component is perturbed at t_o . Hence, the matrix contains stability information integrated over the history of the solution.

In terms of the Green's function, the response of the reference solution at time t , $\underline{\delta}(t)$, to a perturbation of initial conditions at t_o , $\underline{\delta}(t_o)$, is given by

$$(7) \quad \underline{\delta}(t) = \underline{G}(t, t_o, \underline{X}) \underline{\delta}(t_o).$$

An eigenanalysis of the Green's function is performed to assess the growth or shrinkage of $\underline{\delta}$. The matrix \underline{G} is of dimension $N \times N$ and, in general, nonsymmetric. Although the elements of \underline{G} are real, its eigenvalues and eigenvectors may be complex. The Green's function may be expressed in diagonal form

$$(8) \quad \underline{G}(t, t_o, \underline{X}) = \underline{U}^{-1}(t, t_o, \underline{X}) \underline{\Lambda}(t, t_o, \underline{X}) \underline{U}(t, t_o, \underline{X})$$

where \underline{U}^{-1} and \underline{U} are the matrices of left and right eigenvectors. The row vector \underline{U}^{-1} and the column vector \underline{U}_i correspond to λ_i ,

$$(9) \quad \underline{U}^{-1} \underline{G} = \underline{\Lambda} \underline{U}^{-1}$$

and

$$(10) \quad \underline{G} \underline{U}_i = \lambda_i \underline{U}_i.$$

Since G is real, complex eigenvalues may only occur in conjugate pairs. The left and right eigenvectors form a biorthogonal set

$$(11) \quad \underline{U}^{-1} \underline{U} = \underline{1}$$

and G can thus be expressed in terms of these eigenvectors

$$(12) \quad \underline{G} = \underline{U} \underline{\Lambda} \underline{U}^{-1}.$$

The equation for evolution of the perturbation in terms of the eigenvalues and eigenvectors of the Green's function is

$$(13) \quad \underline{\delta}(t) = \underline{U} \underline{\Lambda} \underline{U}^{-1} \underline{\delta}(t_0)$$

The eigenvalues of the Green's function indicate how much the associated modes have grown or diminished in the course of the evolution of the system. The condition for chemical stability is characterized by all eigenvalues less than one in absolute value, and instability by eigenvalues greater than unity in absolute value. A reaction model with an equilibrium state will have a unit eigenvalue indicative of the marginal stability of the equilibrium state.

The eigenvectors form a time dependent coordinate system for the deviations from a solution. The right eigenvectors \underline{U}_i are the modes of evolution for deviations from the time dependent reference solution. The left eigenvectors \underline{U}^{-1} allow for a decomposition of a particular perturbation of initial conditions $\underline{\delta}(t_0)$ into projections along these modes. The inner product of a left eigenvector with the initial perturbation, $\underline{U}^{-1} \cdot \underline{\delta}(t_0)$, is the coefficient of the related right eigenvector which is modulated by the eigenvalue λ_i in the course of evolution. This gives the information needed to adjust initial conditions so as to emphasize or eliminate a particular mode at a later time. Accordingly, from eq. (13) it is evident that projection operators which decompose the evolution of $\underline{\delta}$ into a sum of its independent modes may be defined as $P = \underline{U} \underline{U}^{-1}$.

Negative and positive components of the eigenvectors respectively correspond to concentrations and temperature diminishing and growing from their reference values $\underline{X}(t)$. Furthermore, the eigenvector normalization $\underline{U}^{-1} \underline{U} = \underline{1}$, implied by eq. (11) clearly shows that an arbitrary renormalization of the right eigenvector by a constant C will require a corresponding normalization by $(C)^{-1}$ of the left eigenvector.

Sensitivity Analysis

Sensitivity analysis in the present context is used to determine the role that parameters play in determining stability behavior. Equation (3) may be rewritten as

$$(14) \quad \frac{d\underline{X}}{dt} = \underline{F}(\underline{X}, \underline{\alpha}), \quad \underline{X}(t_0) = \underline{X}_0,$$

to explicitly include the system parameters. The parameters $\underline{\alpha}$ and the initial conditions are assumed to be independent of each other. The Green's function certainly depends on these parameters, i.e., $\underline{G} = \underline{G}[t, t_0, \underline{X}(\underline{\alpha}), \underline{\alpha}]$, where the explicit and implicit dependence on the parameters is indicated.

Consider now the case of a perturbation in the matrix \underline{G} associated with eq. (10). Introducing a linear expansion in \underline{G} , λ_i , and \underline{U}_i yields

$$(15) \quad \underline{G} \longrightarrow \underline{G}(\underline{\alpha}) + \frac{d\underline{G}(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha}$$

$$(16) \quad \lambda_i \longrightarrow \lambda_i(\underline{\alpha}) + \frac{d\lambda_i(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha}$$

$$(17) \quad \underline{U}_i \longrightarrow \underline{U}_i(\underline{\alpha}) + \frac{d\underline{U}_i(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha}$$

where $d\alpha$ is an arbitrary differential change in the vector of parameters. The arbitrariness of $d\alpha$ in eqs. (16) and (17) is predicted on λ_i being nondegenerate. The breakdown of this assumption will be treated as a special case below. Substitution of these relations into eq. (10) gives

(18)

$$\left[\underline{G} + \frac{d\underline{G}}{d\alpha} \cdot d\alpha \right] \left[\underline{U}_i + \frac{d\underline{U}_i}{d\alpha} \cdot d\alpha \right] = \left[\lambda_i + \frac{d\lambda_i}{d\alpha} \cdot d\alpha \right] \left[\underline{U}_i + \frac{d\underline{U}_i}{d\alpha} \cdot d\alpha \right]$$

and collecting terms of like orders in $d\alpha$ produces

(19(a))

$$\underline{G} \underline{U}_i = \lambda_i \underline{U}_i$$

(19(b))

$$[\underline{G} - \lambda_i] \frac{d\underline{U}_i}{d\alpha} \cdot d\alpha = \left[\underline{1} \frac{d\lambda_i}{d\alpha} \cdot d\alpha - \frac{d\underline{G}}{d\alpha} \cdot d\alpha \right] \underline{U}_i$$

Equation (19(a)) is seen to be satisfied automatically since it is exactly the same as eq. (10). In eq. (19(b)), the differential parameter change $d\alpha$ may be treated as arbitrary and thus removed to yield

(20)

$$[\underline{G} - \lambda_i] \frac{d\underline{U}_i}{d\alpha} = \left[\frac{d\lambda_i}{d\alpha} \underline{1} - \frac{d\underline{G}}{d\alpha} \right] \underline{U}_i$$

Multiplying eq. (20) on the left by ${}^i \underline{U}^{-1}$ and utilizing eq. (9) yields

(21)

$$\frac{d\lambda_i}{d\alpha} = {}^i \underline{U}^{-1} \cdot \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_i$$

Returning to eq. (20) and multiplying on the left by ${}^i \underline{U}^{-1}$, $i' \neq i$, the following is obtained

(22)

$$\frac{d\underline{U}_i}{d\alpha} = - \sum_{i' \neq i} \underline{U}_{i'} \left[{}^{i'} \underline{U}^{-1} \cdot \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_i \right] / [\lambda_{i'} - \lambda_i]$$

A similar expression applies to the left eigenvectors.

In the actual application of H_2/O_2 kinetics in this paper, one eigenvalue (say λ_1) is found to dominate all others for virtually all significant times. Then, for the special case of $\lambda_1 \gg \lambda_{i' \neq 1}$, eq. (22) may be rewritten for $i = 1$ in approximate form as

(23)

$$\frac{d\underline{U}_1}{d\alpha} \approx \frac{1}{\lambda_1} \sum_{i' \neq 1} \underline{U}_{i'} \left[{}^{i'} \underline{U}^{-1} \cdot \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_1 \right]$$

Adding and subtracting the quantity

(24)

$$\underline{U}_1 \left[{}^1 \underline{U}^{-1} \cdot \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_1 \right]$$

to the summation yields

(25)

$$\frac{d\underline{U}_1}{d\alpha} \approx \frac{1}{\lambda_1} \sum_{i' \neq 1} \underline{U}_{i'} \left[{}^{i'} \underline{U}^{-1} \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_1 \right] - \underline{U}_1 \left[{}^1 \underline{U}^{-1} \cdot \frac{d\underline{G}}{d\alpha} \cdot \underline{U}_1 \right]$$

where the summation is now over all i' . Making use of orthonormality, $\sum_i \underline{U}_i {}^i \underline{U}^{-1} = \underline{1}$, and eq. (21) yields,

(26)

$$\frac{d\underline{U}_1}{d\alpha} = \frac{1}{\lambda_1} \left[\frac{d\underline{G}}{d\alpha} \underline{U}_1 - \underline{U}_1 \frac{d\lambda_1}{d\alpha} \right] = \frac{1}{\lambda_1} \left[\frac{d\underline{G}}{d\alpha} - \underline{1} \frac{d\lambda_1}{d\alpha} \right] \underline{U}_1$$

Under the same condition $\lambda_1 \gg \lambda_{i \neq 1}$, note also from eq. (22) that $d\bar{U}_i/d\alpha_i$, $i \neq 1$, will have essentially no component along \bar{U}_1 .

The eigensensitivity calculations provide further information about the dynamical behavior of a particular model under consideration. Eigenvalue sensitivities are indicative of the effect on system stability of local excursions in the vicinity of the parameter space operating point. The magnitude and sign of the sensitivities provide a measure of whether changes in the system will increase or decrease stability. Eigenvector sensitivities yield information on how the dynamical modes of evolution are affected by alterations in the system. Particular combinations of state space variables may act together upon parameter variation, and this information is conveniently summarized in the components of the eigenvector sensitivities. Again, the magnitude and signs of these components provide this quantitative information.

The above eigensensitivity analysis assumes that the system is nondegenerate. For the analysis of H_2/O_2 kinetics in this article, this assumption was valid for the largest eigenvalue whenever it dominated all others, which was for virtually all significant reaction times. However, in many problems, degeneracy may be important and the necessary modifications to the above equations for the degenerate case are presented for completeness in Appendix A. A potentially important case not included in this analysis arises for near degeneracy where the purely degenerate or nondegenerate forms are not strictly valid. In the present work, the Green's function, G_{ij} , and the parametric sensitivities of G_{ij} , $\partial G_{ij}/\partial \ln k_i$, are obtained using the AIM computer code [31].

Comparison to Variational Equation Stability Analysis

The traditional (variational equations) approach to stability analysis entails an eigenanalysis of the Jacobian \underline{J} . The first variational equation for the system of eq. (3) is

$$(27) \quad \dot{\underline{\delta}}(t) = \underline{J}(\underline{X}) \underline{\delta}(t)$$

When $\underline{J}(\underline{X})$ can be assumed constant (e.g., for small time intervals from the initial time t_0), the integrated equation of motion for $\underline{\delta}$ is

$$(28) \quad \underline{\delta}(t) = \exp[\underline{J} \cdot (t - t_0)] \underline{\delta}(t_0),$$

which can be compared to eq. (7). The eigenvalues of the Jacobian \underline{J} prescribe how perturbations of the initial condition behave for small time intervals near t_0 . Growth of $\underline{\delta}$ is indicated by the matrix $\exp[\underline{J} \cdot (t - t_0)]$ having an eigenvalue greater than unity in absolute value. If \underline{J} has an eigenvalue with a positive real part, this condition holds and instability is indicated. Except for autonomous linear systems, eq. (28) should be thought of as only being valid near the initial condition $\underline{X}(t_0)$.

The variational equation analysis is considered to be local in two ways: first, it depends on the position in state space of the solution, and second, through the assumption that $\underline{J}(\underline{X})$ is constant, it is only valid for times near the point in time where the eigenvalues of $\underline{J}[\underline{X}(t)]$ are calculated. Objections to this approach are based on the following possibilities; although a nearby solution may be diverging from the reference solution at some point, it may

later converge to it. An analysis of stability based on the Jacobian alone does not incorporate this possibility in a useful way.

Note that if the Jacobian is independent of time, eq. (5) may be integrated to yield

$$\underline{G}(t, t_0, \underline{X}) = \exp[\underline{J} \cdot (t - t_0)],$$

in which case the two methods coincide. This furthermore points out the fact that the variational equations approach is local in time. The Green's function analysis is local in the sense that $|\delta|$ is assumed to remain small over the course of its evolution. However, there is no restriction that t remain small. Hence, if solutions near the reference solution diverge from it in some time interval, but converge to it in others, then the net effect is still incorporated in the Green's function.

Results

Figure 1 is a plot of the classical explosion limits for a stoichiometric mixture of hydrogen and oxygen (from Lewis and VonElbe [32]). The three

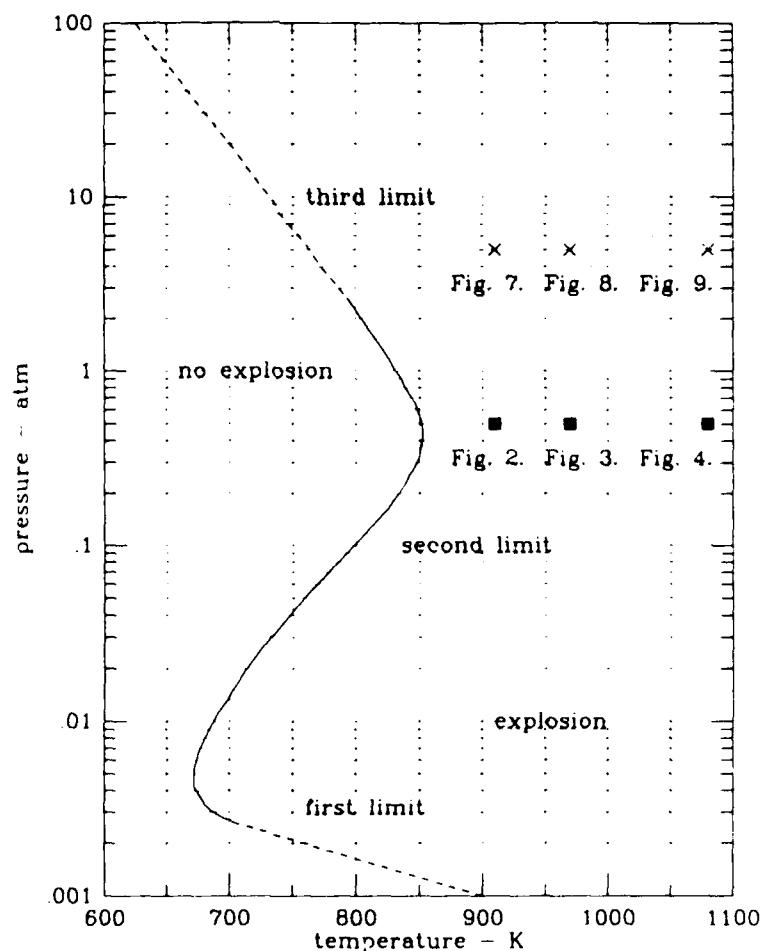


Figure 1. Explosion limits for a stoichiometric mixture of hydrogen and oxygen (from [32]). The dashed lines are extrapolations of the first and third limits. The symbols (crosses and squares) denote the initial temperature and pressure conditions of the kinetic calculations described in Figures 2-11.

limits have been the subject of numerous articles (e.g., [33–36]), most recently by Maas and Warnatz [37] who predicted the three limits by modeling the detailed kinetics in spherical vessels with time-dependent, 1-D spatial calculations.

The present work has concentrated on the ignition characteristics of explosive mixtures only. A dilute stoichiometric mixture of 1% hydrogen and 0.5% oxygen reacting in nitrogen was considered. The dilute mixture was chosen in order to limit the total heat release to a temperature rise of approximately 100 K. Figures 2–4 present the kinetics and stability analysis results for three computational experiments, all with an initial pressure of 0.5 atm and with initial temperatures of 910 K, 970 K, and 1080 K, respectively. The location of the initial conditions are illustrated on the pressure-temperature phase-plane of Figure 1. (Note that the classical explosion limits of dilute stoichiometric mixtures in nitrogen do not necessarily coincide with the limits of nondilute mixtures.) Since the mixture was dilute, the trajectory of the kinetics through pressure-temperature phase space follows a nearly constant pressure line to a final temperature approximately 100 K higher than T_0 .

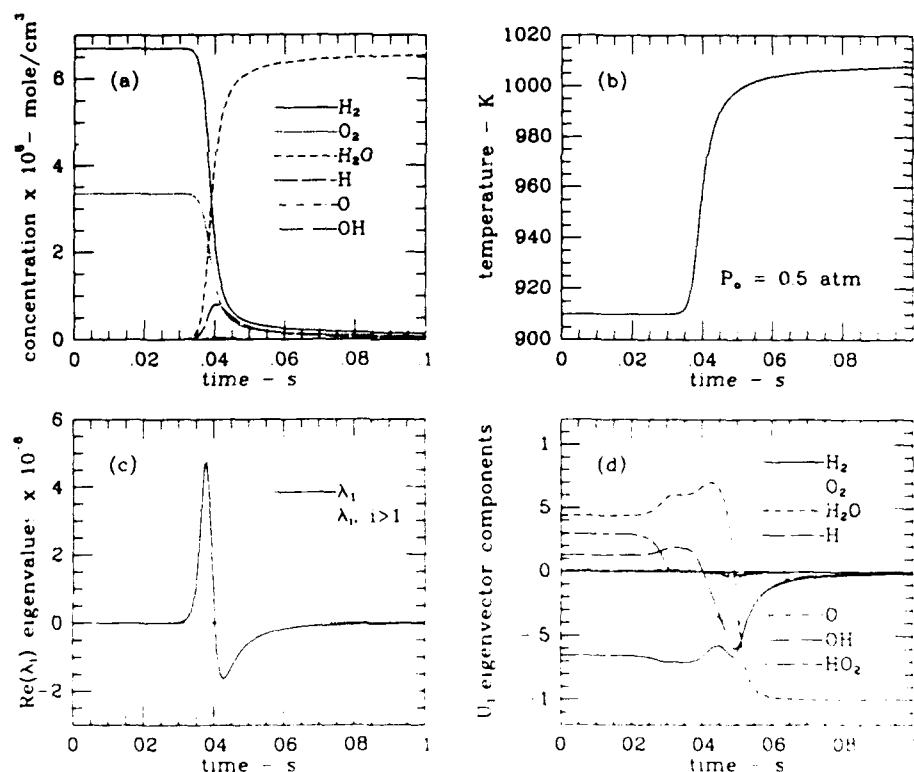


Figure 2. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 910$ K, $P = 0.5$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 0.5 atm and 1010 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

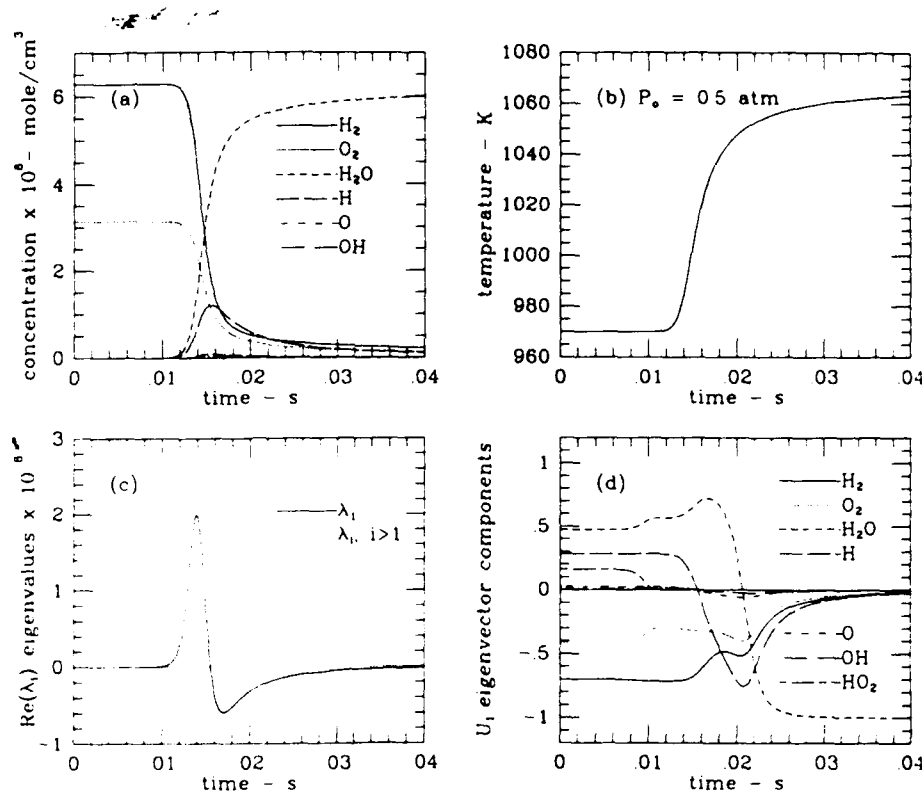


Figure 3. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 970$ K, $P = 0.5$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 0.5 atm and 1070 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

The species concentration and temperature profiles are all similar indicating an increase in reaction rate with increasing temperature, and thus, shorter induction and reaction times (see parts (a) and (b) of each figure). Also, the higher the initial temperature, the higher the H, O, and OH radical concentrations.

In part (c) of each figure, the real parts of the largest and remaining other eigenvalues are given. Note that the reaction dynamics of each system are controlled by a single eigenvalue (λ_1), and that the magnitudes of the real part of this eigenvalue are extremely large (of the order 10^8), and hence, the mixtures are highly explosive. As the temperature is increased, the maximum magnitudes of λ_1 are observed to decrease. (More will be said on the strengths of the explosions later). The magnitudes of the real part of the remaining eigenvalues were generally less than unity, except for a few unique reaction times. In particular, $\text{Re}(\lambda_2)$, the second largest eigenvalue exceeded unity and equaled $\text{Re}(\lambda_1)$ near the location where the two eigenvalues are indicated to cross in the figures; for example at $t = 0.04$ s for the mixture with $T_0 = 910$ K (see Fig. 2(c)). Further, the imaginary components for both λ_1 and λ_2 were zero, except when $\text{Re}(\lambda_2)$ equaled $\text{Re}(\lambda_1)$. During this period,

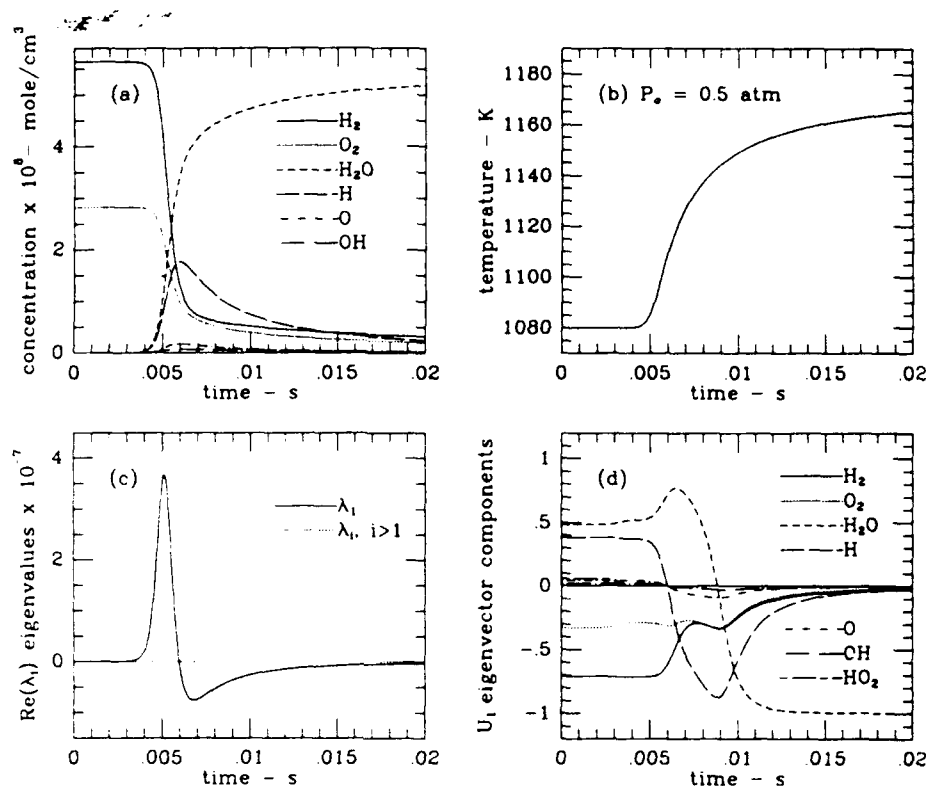
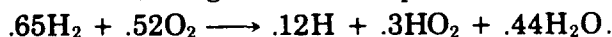


Figure 4. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 1080$ K, $P = 0.5$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 0.5 atm and 1180 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

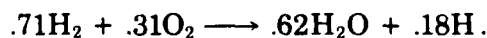
the two eigenvalues are complex conjugates. Hence, $|\text{Im}(\lambda_1)| \ll |\text{Re}(\lambda_1)|$ for all reaction times of interest here and thus only the real part of λ , is plotted.

Comparison of the species concentration and temperature profiles with the profile for $\text{Re}(\lambda_1)$ enables the induction time to be defined as the time from t equal zero to the first maxima in the eigenvalue profile. Due to the dominance of λ_1 , the subsequent analysis will focus on it and its associated eigenvector \underline{U}_1 .

The eigenvalues describe the magnitude of change in species concentrations and temperature. The associated right eigenvectors specify the direction of change. The components of the eigenvector associated with the λ_1 are reported in part (d) of each figure. Here, the components of \underline{U}_1 were normalized according to $U_{1,i}/(\sum_i U_{1,i}^2)^{1/2}$ where the summation excludes the component corresponding to the temperature variable. During the induction period for the mixture with $T_0 = 910$ K, the relative change in species concentrations can be characterized by two distinct overall stoichiometric vectors. For $0 < t < 0.025$ s, the growth of the perturbation follows



The eigenvector then rotates to a new direction with constant components for $0.03\text{s} < t < 0.036$ s with an overall stoichiometric vector of



With increasing initial temperature, the distinction between the two overall reactions during the induction period disappears. Inspection of the eigenvector components reveals loss of HO_2 formation as the temperature is increased (compare Figs. 2(d), 3(d), and 4(d)). Once appreciable consumption of the initial reactants begin, the eigenvector again rotates and continues to until the reaction nears completion. For the mixture with $T_0 = 910$ K, examples of overall stoichiometric vectors are $.68\text{H}_2 + .3\text{O}_2 \rightarrow .6\text{H}_2\text{O} + .16\text{H}$ at 25% H_2 consumption, $.66\text{H}_2 + .33\text{O}_2 \rightarrow .66\text{H}_2\text{O}$ at 50% H_2 consumption, and $.6\text{H}_2 + .35\text{O}_2 + .2\text{H} \rightarrow .7\text{H}_2\text{O}$ at 75% H_2 consumption. Hence, the direction of the eigenvector is never constant during the consumption of major reactants. Comparison of \underline{U}_1 for different initial temperatures shows that the corresponding eigenvector components to be nearly the same during the first half of the reaction, and that the H-atom component increases and the H_2 component decreases during the latter half of the reaction as the initial temperature is increased.

Note that near the end of H_2 consumption, the response of the system is entirely in the direction of H_2O formation. This is to be expected because at large reaction times, water vapor is the favored thermodynamic product. During the early period of reaction, it was also generally observed that a nearly identical reaction vector could be obtained if the stoichiometric coefficients associated with the elementary reactions which had the largest fluxes were each scaled by their corresponding fluxes and then summed.

The sensitivities of λ_1 to the elementary rate constants of the mechanism are given in Figure 5. At 910 K, λ_1 is sensitive to the rate constants of $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$ and $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$. Other rate constants have a relatively small sensitivity. As the temperature is increased, the maximum magnitudes of both the absolute ($\partial\lambda_1/\partial\ln k_j$) and relative ($\partial\ln\lambda_1/\partial\ln k_j$) sensitivity gradients decrease. Further, the sensitivity of λ_1 to the rate constant of $\text{H} + \text{O}_2 + \text{M}$ is reduced significantly with increasing temperature compared to that of the branching reaction. This trend is in agreement with the loss of the HO_2 component of \underline{U}_1 at high temperatures. Reactions of secondary importance include $\text{H}_2 + \text{O} \rightarrow \text{OH} + \text{O}$, $\text{H}_2 + \text{O}_2 \rightarrow \text{HO}_2 + \text{H}$, $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$, $\text{HO}_2 + \text{H} \rightarrow 2\text{OH}$, $\text{HO}_2 + \text{H} \rightarrow \text{H}_2 + \text{O}_2$, and $\text{OH} + \text{O} \rightarrow \text{H} + \text{O}_2$, in decreasing order of importance. Note that since the system is controlled by a single eigenvalue, the reactions discussed above are ranked with respect to the entire system and not with respect to a single dependent variable, as are the elementary sensitivity gradients, $\partial X_i/\partial\ln k_j$, for different choices of X_i .

The sensitivity of the eigenvector direction to the elementary rate constants of the mechanism is illustrated in Figure 6 for the system with an initial temperature of 910 K. The sensitivity gradients, $\partial U_{1,i}/\partial\ln k_j$, for H_2 , O_2 , H_2O , H , HO_2 , and OH components are shown. To evaluate these gradients, the approximation $\lambda_1 \gg \lambda_{i \neq 1}$ was made, allowing for use of eq. (26), which is valid for all time shown in Figure 2 except near 0.04s where λ_1 passes through zero. The important reactions are the same as those found important to λ_1 ; however, the order of ranking of important reactions was not always identical for each species. The most sensitive species are the H_2O , H_2 , O_2 , HO_2 , and H components, listed in decreasing order of sensitivity. However, examination of the corresponding normalized sensitivities for H_2 , O_2 , H_2O , and HO_2 shows that during the interval $0 < t < 0.025\text{s}$, the

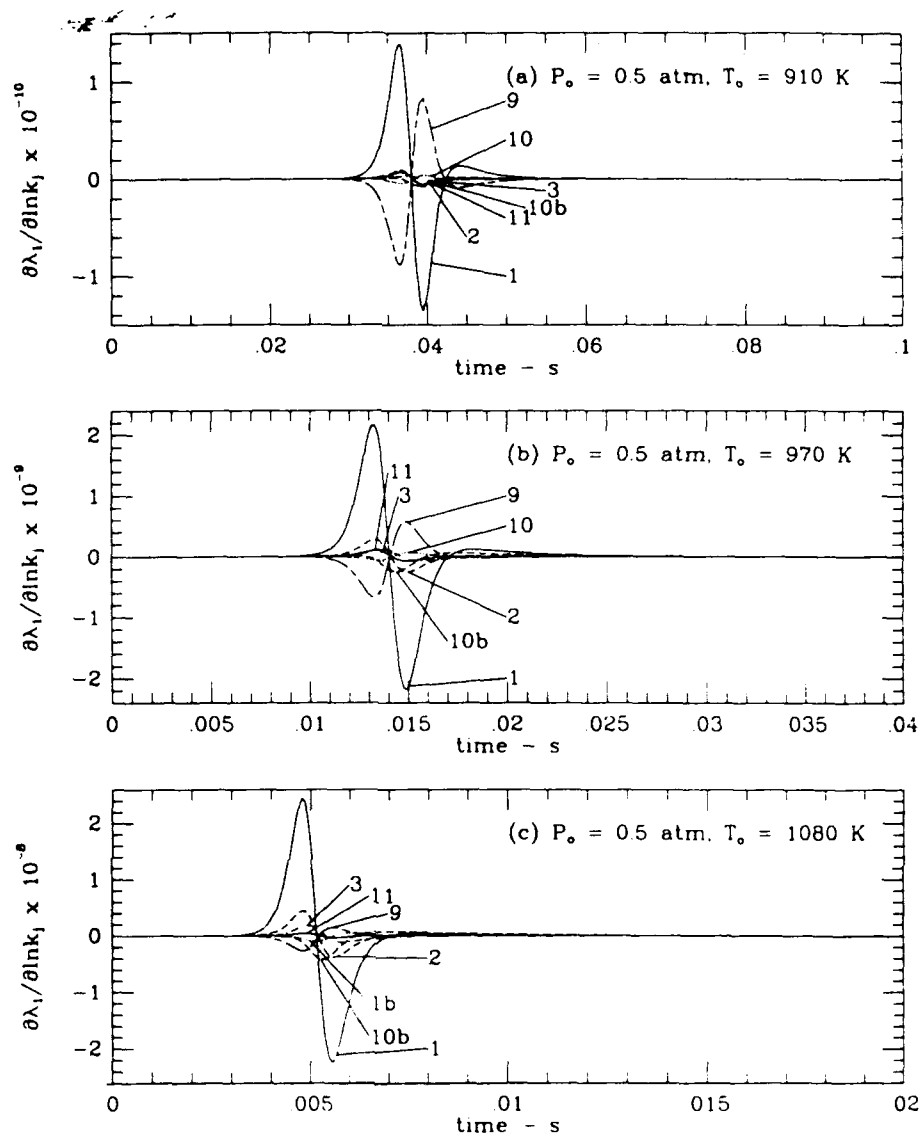


Figure 5. Sensitivity gradients of the largest eigenvalue with respect to various reaction rate constants. Initial conditions: $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, $P = 0.5$ atm, (a) $T = 910$ K, (b) $T = 970$ K, (c) $T = 1080$ K. The numbers denote the reactions of Table II. The letter "b" after the number denotes the backward reaction.

relative responses of these species to perturbations in k_j are all approximately equal, with the signs of the gradients for reactants, H_2 and O_2 , opposite to those of products, H_2O and HO_2 . Hence, if either k_1 or k_9 is perturbed, the species coefficients are observed to change dramatically, but in a manner such that the direction of the reaction vector changes little, except in the formation of H-atoms. From Figure 6(d), an increase in either k_1 or k_9 produces a slight increase in the amount of H-atom formation. For times greater than 0.03s, the HO_2 component becomes insensitive to perturbations in any of the rate constants. Although both the H-atom and the OH radical components are relatively insensitive to rate constant perturbations, it is interesting to note that the H-atom is sensitive to reaction 2 while the OH radical is sensitive to reaction 3. At higher temperatures (970 K and

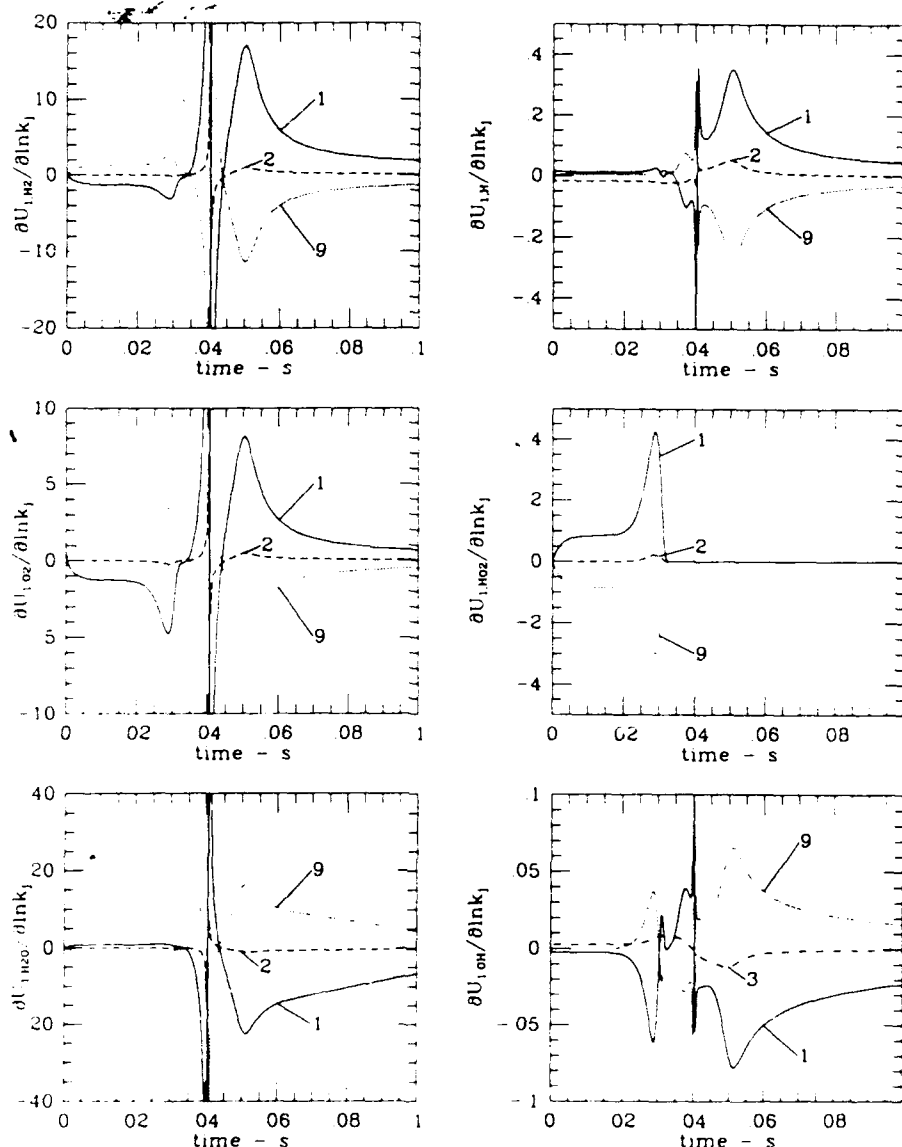


Figure 6. Sensitivity gradients of selected eigenvector components associated with the largest eigenvalue with respect to various reaction rate constants. Initial conditions: $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, $P = 0.5$ atm, $T = 910$ K. The numbers denote the reactions of Table II. The letter "b" after the number denotes the backward reaction.

1080 K), the sensitivity gradients of eigenvector components were found consistent with those observed at 910 K.

In comparison to the results at 0.5 atm, Figures 7–9 present the kinetic and stability analysis results for another three computational experiments, again with the same initial temperatures of 910 K, 970 K, and 1080 K, but all with an initial pressure of 5 atm. The location of the initial conditions are also illustrated on the pressure-temperature phase-plane of Figure 1.

Again, all three systems are controlled by a single eigenvalue. However, at low temperatures, the ignition process is characterized by an eigenvalue with a low magnitude (order of 10^4 , see Fig. 7) compared to that at high tem-

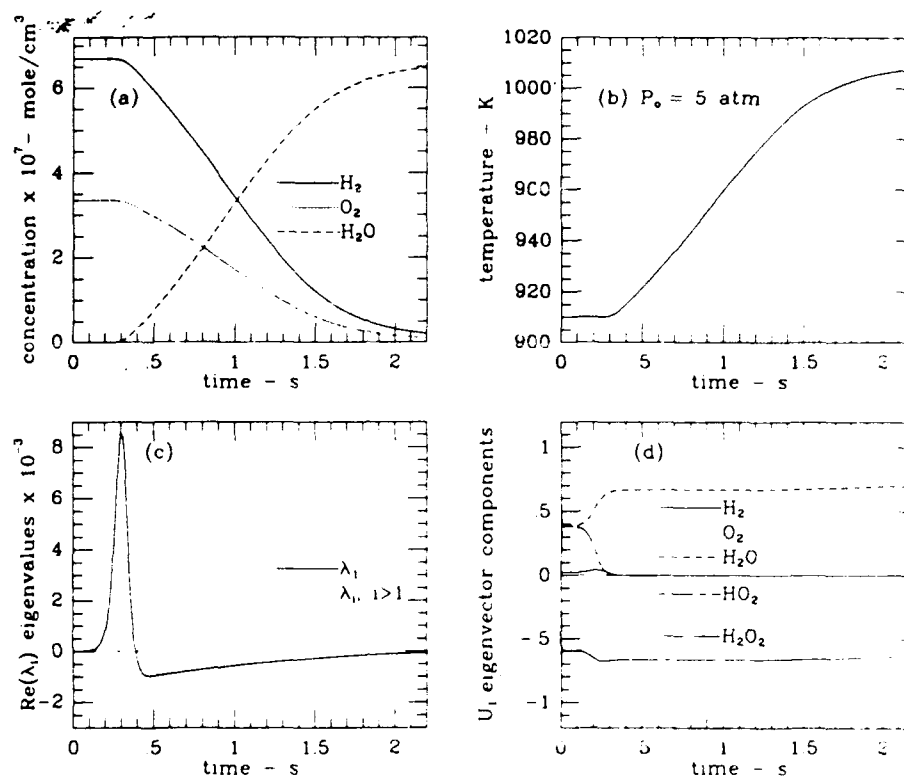


Figure 7. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 910$ K, $P = 0.5$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 5.0 atm and 1010 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

peratures where the magnitude (order of 10^7 , see Fig. 9) is close to those observed at 0.5 atm. Based on these magnitudes, the low temperature system can be classified as "weak" ignition while the high temperature system can be classified as "strong" ignition, as discussed earlier. Note that ignition at 0.5 atm was all "strong" ignition. The transition from "weak" to "strong" ignition is clearly illustrated for the intermediate temperature system of Figure 8. The eigenvalue first shows "weak" ignition at about 0.08s and then "strong" ignition at about 0.116s. Transition occurs at a temperature of 1028 K.

At 910 K, \underline{U}_1 during the induction period is $.6\text{H}_2 + .6\text{O}_2 \rightarrow .37\text{H}_2\text{O} + .37\text{HO}_2 + .045\text{H}_2\text{O}_2$. Note that under the conditions of "weak" ignition, the eigenvector \underline{U}_1 remains constant during the consumption of major reactants with an overall stoichiometric vector of $.66\text{H}_2 + .33\text{O}_2 \rightarrow .66\text{H}_2\text{O}$.

At 970 K, \underline{U}_1 has nearly the same components during the induction time and first stage of reaction as found at 910 K. However, when the temperature of the mixture reaches 1028 K, the eigenvector begins to rotate as observed at 1080 K and for all three temperatures at 0.5 atm.

The "weak" ignition process is sensitive to rate constants of a different group of reactions (see Fig. 10(a)). In decreasing order of importance, these

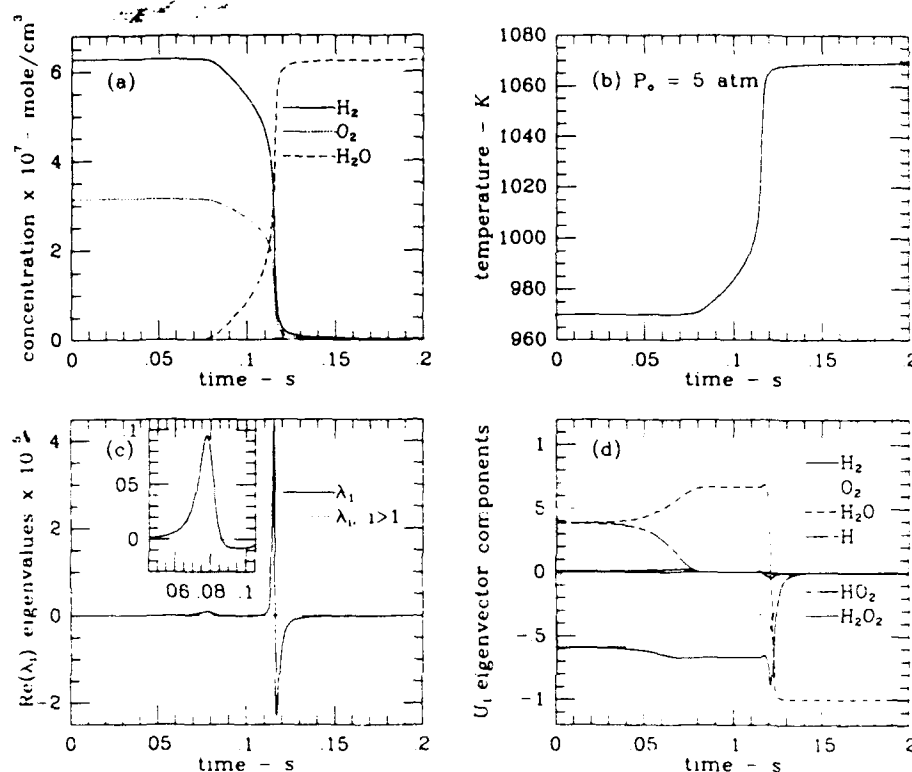


Figure 8. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 970$ K, $P = 5.0$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 5.0 atm and 1070 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

reactions are, $\text{H}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{H}$, $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$, $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$, $\text{H}_2\text{O}_2 + \text{M} \rightarrow \text{OH} + \text{OH} + \text{M}$, $\text{H}_2 + \text{O}_2 \rightarrow \text{H} + \text{HO}_2$, $\text{H} + \text{HO}_2 \rightarrow \text{H}_2 + \text{O}_2$, $\text{H} + \text{HO}_2 \rightarrow \text{OH} + \text{OH}$ and $\text{HO}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{O}_2$. During the consumption of major reactants, the same reactions remain important; however, the order of ranking changes. For example, at 50% consumption H_2 , the ordering of most important reaction rate constants is $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$, $\text{H} + \text{HO}_2 \rightarrow \text{H}_2 + \text{O}_2$, $\text{H}_2 + \text{O}_2 \rightarrow \text{H} + \text{HO}_2$, $\text{HO}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{O}_2$, $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$, $\text{H} + \text{HO}_2 \rightarrow \text{OH} + \text{OH}$, $\text{H}_2\text{O}_2 + \text{M} \rightarrow \text{OH} + \text{OH} + \text{M}$, $\text{H}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{H}$, and $\text{HO}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{O}_2$. The "strong" ignition process at 5 atm (Fig. 10(c)) is sensitive to the same reactions as the "strong" ignition process at 0.5 atm. As might be expected, the first stage of the intermediate temperature (Fig. 10(b)) ignition process is sensitive to the rate constants of reactions characteristic of "weak" ignition while the second ignition process is sensitive to the reactions important to "strong" ignition.

The sensitivity of the H_2 , O_2 , H_2O , HO_2 , and H_2O_2 eigenvector components at $T = 910$ K are presented in Figure 11. The condition of $\lambda_1 \gg \lambda_{i \neq 1}$, allowing for use of eq. (26), is satisfied everywhere except near $t = 0.4$ s. At $t = 2.2$ s, λ_1 is still approximately 100 times larger than the next largest

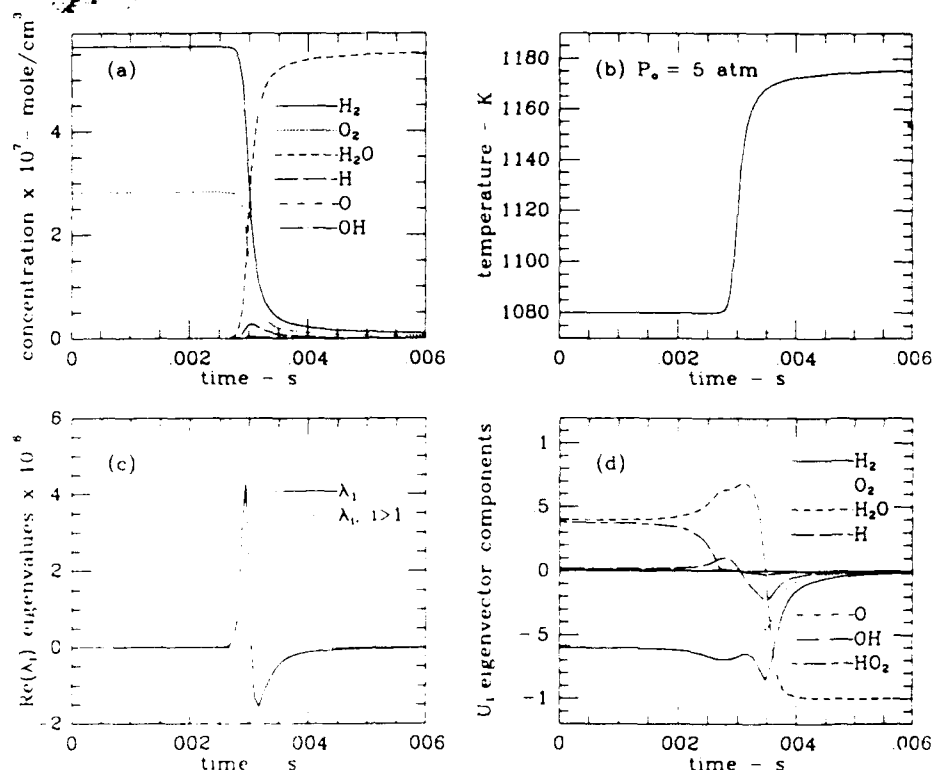


Figure 9. Kinetic and stability analysis results for a dilute stoichiometric mixture of hydrogen and oxygen reacting in a constant volume adiabatic bath of nitrogen. Initial conditions: $T = 1080$ K, $P = 5.0$ atm, $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, (a) species concentrations, (b) temperature. Note the temperature rise of approximately 100 K. Since the mixture is dilute, the pressure remains nearly constant, and hence, the trajectory of the kinetics on the pressure-temperature phase-plane of Figure 1 is approximately a horizontal line ending at 5.0 atm and 1180 K, (c) the real parts of the eigenvalues, (d) the components of the eigenvector associated with the largest eigenvalue.

eigenvalue. Again the most sensitive components are the H_2 , O_2 , H_2O , and HO_2 species. However, relative to the results at 0.5 atm, the stable species are about an order of magnitude more sensitive while the unstable species are about an order of magnitude less sensitive. At 50% consumption H_2 , the ranking of important reactions follows the order: $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$, $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$, $\text{H} + \text{HO}_2 \rightarrow 2\text{OH}$, $\text{H} + \text{HO}_2 \rightarrow \text{H}_2 + \text{O}_2$, $\text{H}_2\text{O}_2 + \text{M} \rightarrow \text{OH} + \text{OH} + \text{M}$, $\text{H}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{H}$, $\text{HO}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{O}_2$, $\text{HO}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{O}_2$, $\text{HO}_2 + \text{O} \rightarrow \text{O}_2 + \text{OH}$, $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$ and $\text{H}_2 + \text{O} \rightarrow \text{OH} + \text{H}$.

Using the same dilute mixture as analyzed at 0.5 and 5 atm, the temperatures of transition from weak to strong reaction were evaluated from kinetic calculations for pressures ranging from 1 to 10 atm. The results, plotted as solid triangles in Figure 12, show transition to occur over a range of temperatures, which is wider at lower pressures than at high pressures. This range of transition temperatures resulted from varying the initial temperature of the mixture over approximately 20 K. For example, at a pressure of 6 atm, the resulting variation in transition temperature, defined here as the temperature corresponding to the second positive peak in the maximum eigenvalue profile, was 4 K for a variation in T_0 of 20 K.

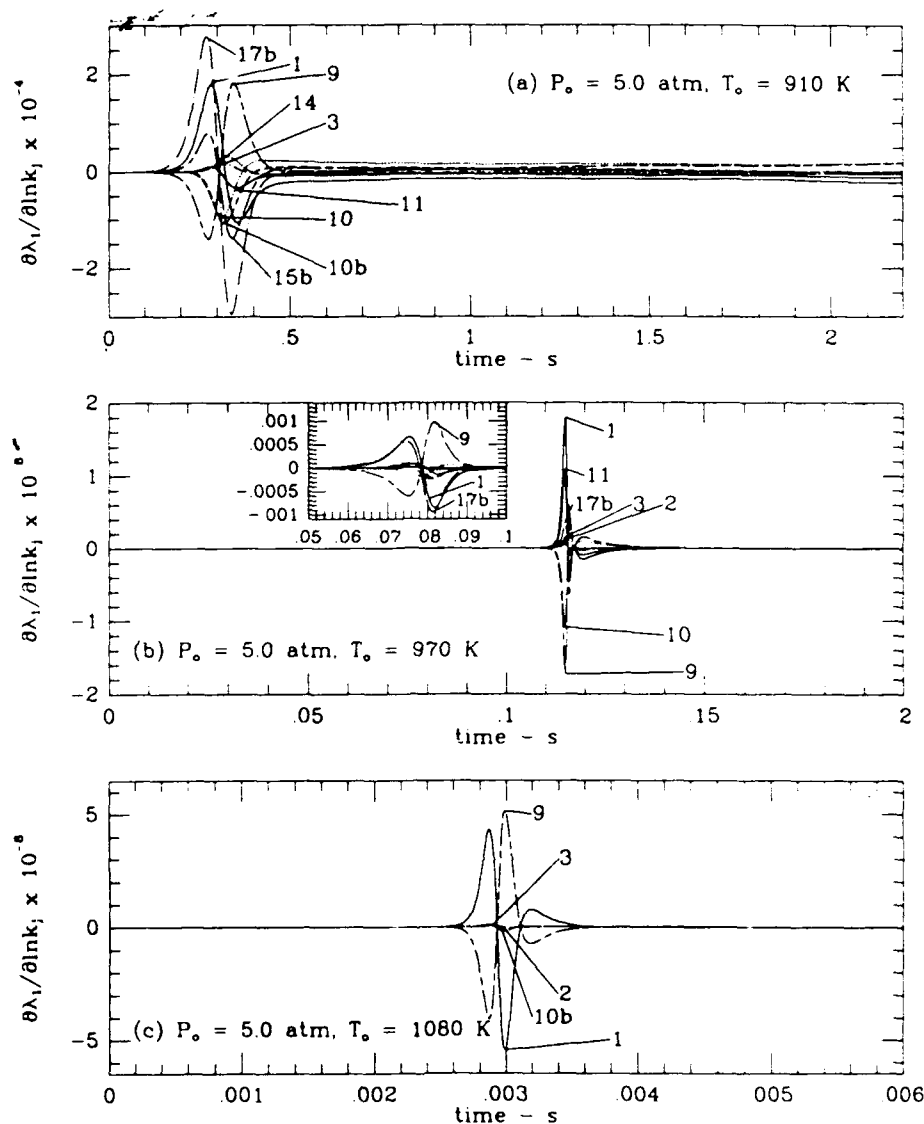


Figure 10. Sensitivity gradients of the largest eigenvalue with respect to various reaction rate constants. Initial conditions: $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, $P = 5.0$ atm, (a) $T = 910$ K, (b) $T = 970$ K, (c) $T = 1080$ K. The numbers denote the reactions of Table II. The letter "b" after the number denotes the backward reaction.

The classical extended second limit, evaluated from the relationship $[M] = 2k_1/k_9$ [32], is also plotted in Figure 12. For a given pressure, transition is observed to occur at a lower temperature than indicated by the classical extended second limit. The deviation appears to widen as the pressure is increased. According to the sensitivity analysis results of Figure 10(b), this deviation may result from neglecting the effects of reactions 17(b), 10, and 11 in the derivation of the classical second limit.

Note that in the explosive region above the "extended" second limit and the third limit, formation of HO_2 and H_2O_2 and their consumption are important to the rate of reaction. The hydroperoxy radical is formed almost entirely through $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$. Consumption of HO_2 occurs through reaction with H-atoms, $\text{HO}_2 + \text{H} \rightarrow \text{OH} + \text{OH}$ and $\text{HO}_2 +$

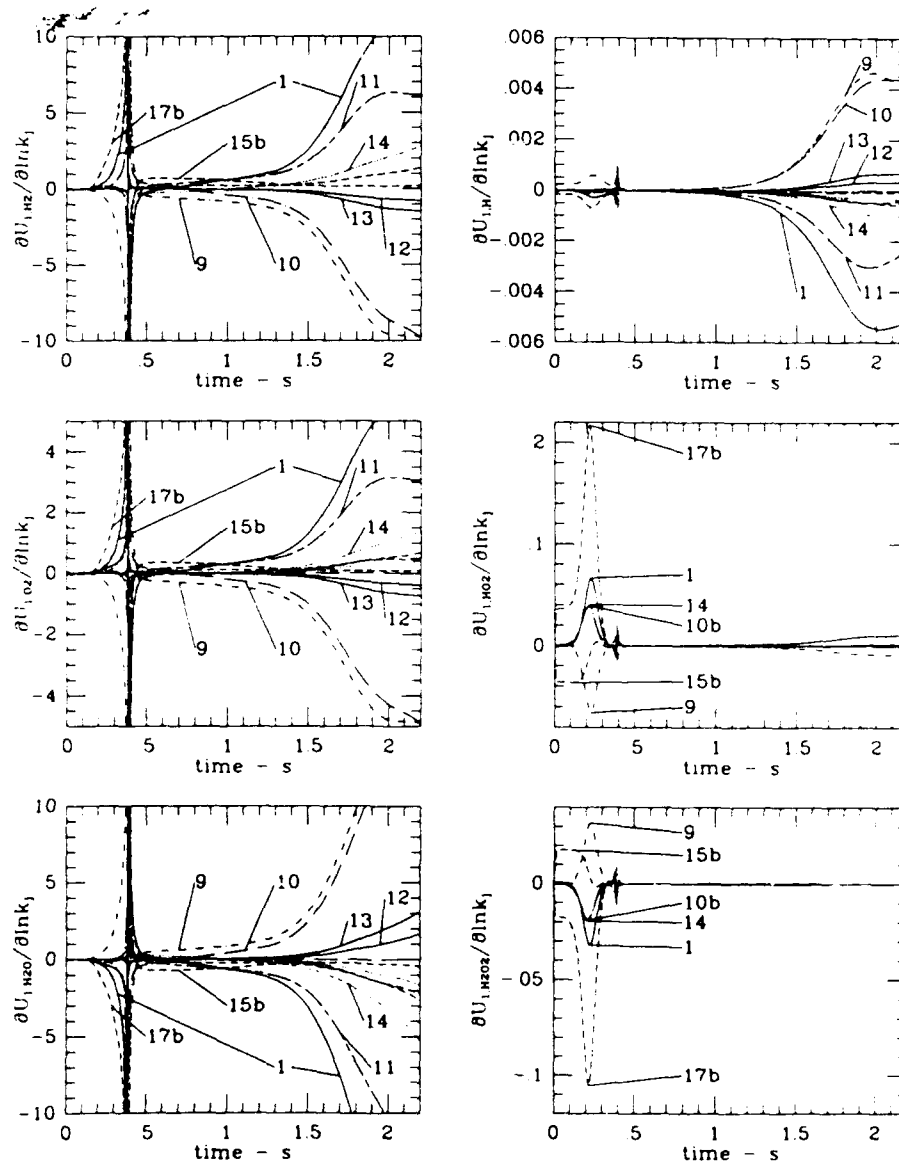


Figure 11. Sensitivity gradients of selected eigenvector components associated with the largest eigenvalue with respect to various reaction rate constants. Initial conditions: $X(\text{H}_2) = 0.01$, $X(\text{O}_2) = 0.005$, $X(\text{N}_2) = 0.985$, $P = 5$ atm, $T = 910$ K. The numbers denote the reactions of Table II. The letter "b" after the number denotes the backward reaction.

$\text{H} \rightarrow \text{H}_2 + \text{O}_2$, or with another HO_2 , $\text{HO}_2 + \text{HO}_2 \rightarrow \text{H}_2\text{O}_2 + \text{O}_2$. The first of these steps is chain propagating while the latter two are terminating. Hydrogen peroxide is formed either by the self reaction of HO_2 or by reaction of HO_2 with H_2 , $\text{HO}_2 + \text{H}_2 \rightarrow \text{H}_2\text{O}_2 + \text{H}$. Consumption of H_2O_2 is by dissociation, $\text{H}_2\text{O}_2 + \text{M} \rightarrow 2\text{OH} + \text{M}$. Almost all of the H_2O is formed via $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$. Neglecting the small amount of branching due to $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$ in this pressure-temperature region, the only chain sequence which leads to chain branching is formation of H_2O_2 by reaction of HO_2 with H_2 followed by thermal decomposition of H_2O_2 . This sequence is slow relative to other chain propagating steps and hence the overall reaction is nearly straight chain, which is also evident from the absence of radical

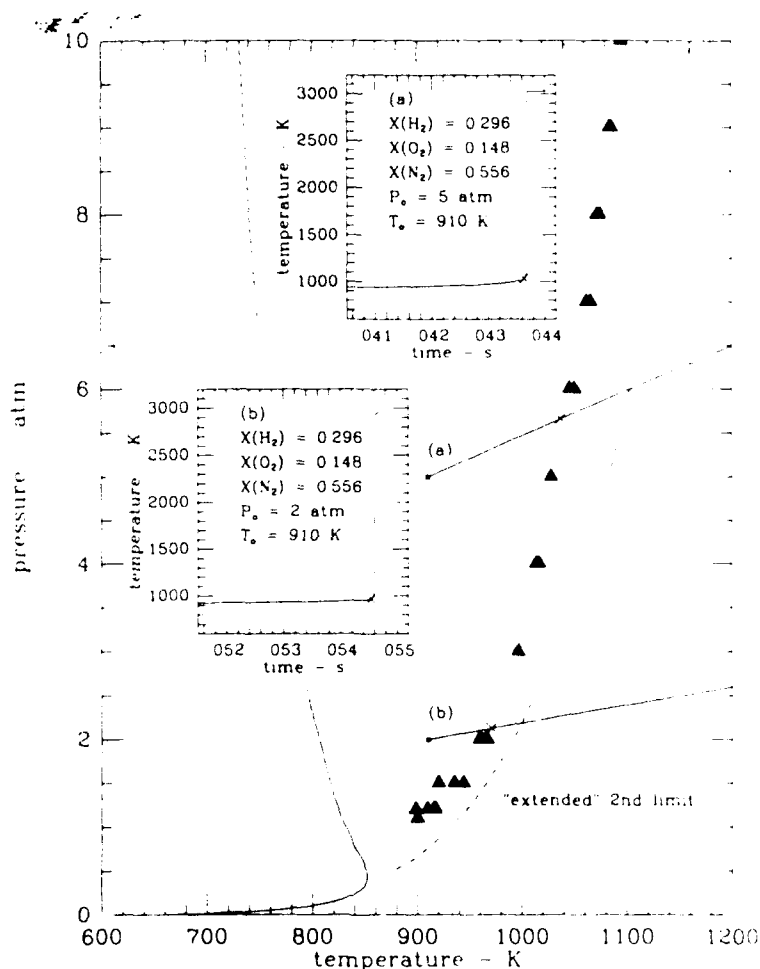
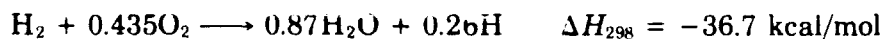
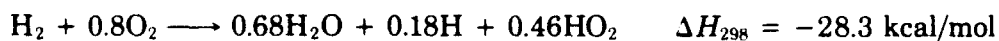
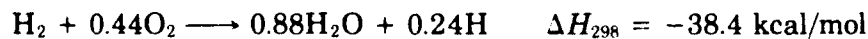


Figure 12. Explosion limits for stoichiometric mixtures of hydrogen and oxygen. The solid line in the lower left hand corner of the figure and the dashed line are the second and third explosion limits shown earlier in Figure 1. The dash-dot-dash line is the classical "extended" second limit. The solid triangles are the transition temperatures calculated from the eigenanalysis of the kinetic solutions for the dilute mixture consisting of 1% H_2 , 0.5% O_2 , and 98.5% N_2 . The two insert figures report the temperature profiles for a stoichiometric H_2 /air mixture with an initial temperature of 910 K and initial pressures of 5 atm (insert a) and 2 atm (insert b). The x's denote the temperatures where d^2T/dt^2 was a maximum for the two nondilute calculations.

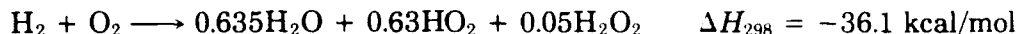
species in the eigenvector components of Figure 7 during H_2 and O_2 consumption. However, note from the overall reaction vectors that the induction reaction and the reaction which occurs during the first 50% consumption of H_2 are more exothermic above the extended second limit than below this limit. For example, the overall reactions and associated exothermicity for consumption of one mol of H_2 at 910 K and 0.5 atm were:



during the induction period and



during the first 50% consumption of H_2 . At 910 K and 5 atm, the corresponding results were



during the induction period and



during the consumption of H_2 . This result is of particular importance to nondilute mixtures.

For nondilute mixtures, e.g., a stoichiometric H_2 /air mixture, a similar eigenanalysis does not produce the dual character in the maximum eigenvalue as observed in Figure 8. Instead, the maximum eigenvalue after a short period of time grows rapidly to a value of ca. 10^3 , and then continues to grow monotonically and more slowly up to ca. 10^4 rather than decrease as observed for the dilute mixture at $T = 970$ K and $P = 5$ atm. A sudden exponential growth to a value of ca. 10^{11} is then observed. After this peak, λ_1 goes negative reaching a minimum peak. During the growth of the eigenvalue from 10^3 to 10^4 , the mixture heats up appreciably due to the exothermicity of the HO_2 reactions until it reaches a temperature near the extended second limit, at which point the eigenvalue rapidly jumps to 10^{11} . For nondilute mixtures, this temperature essentially represents the ignition temperature, with the chemistry prior to this temperature representative of induction chemistry, i.e., continued growth in the radical pool. In Figure 12, both the temperature profiles as a function of reaction time (insert figures) and the pressure-temperature trajectories (solid lines) are reported for this stoichiometric H_2 /air mixture with an initial temperature of 910 K and initial pressures of 5 atm (insert (a)) and 2 atm (insert (b)). The "x's" on the temperature-time profiles and the pressure-temperature trajectories correspond to the temperatures where d^2T/dt^2 equalled a maximum. It is apparent that between the extended second limit and third limit, the overall reaction is characterized by a thermal explosion until the transition temperature is reached where the reaction becomes a branched chain explosion.

Finally, according to the results of Figure 12, the transition temperature for the dilute stoichiometric H_2/O_2 mixture reacting in N_2 at 1 atm is approximately 910 K while at 5 atm this temperature increases to approximately 1028 K. For dilute mixtures in Ar, the transition temperatures shift to slightly lower values because of the decrease in efficiency of Ar as a third body in the recombination reaction $H + O_2 + M \rightarrow HO_2 + M$. Assuming that the effect of mixture stoichiometry on the transition temperatures is small, all of the experimental ignition delays reported here at 1 atm (Table IV) and only one at 5 atm ($T = 1075$ K, Table III) can be classified as "strong" ignition. The remainder of the ignition delay data at 5 atm is likely controlled by "weak" ignition. This separation of the data indicates that the agreement between experiment and model is consistently better for "strong" ignition than for "weak" ignition. According to the sensitivity analysis results, model parameters which should be considered for possible refinement to improve the agreement of weak ignition include the rate constants of reactions 17(b), 9, 15(b), 10, and 10(b), and also the heat of formation for HO_2 . The heat of formation of HO_2 is important because in the model the rate constants for reactions 17(b) and 15(b) were obtained from

the forward rate constants and thermochemical data. Of all the thermochemical data necessary for these reactions, the data for HO_2 have the greatest uncertainties.

Clearly, more accurate and detailed experimental data are needed before further model validation and refinement can be made. In fact, the present analysis indicates that the maxima in OH concentration may be a poor measure of experimental ignition delays for the low temperature experiments of Skinner and Ringrose. By the time the OH concentration has reached its maximum, significant amounts of H_2 have been consumed and heat released. Hence, the temperature history has to be well characterized, since the reaction is no longer isothermal.

Conclusions

In the present article, the extended second limit is shown to be a kinetic boundary important to both the ignition and reaction characteristics of dilute H_2/O_2 mixtures. Transition is generally observed to occur at temperatures lower than predicted by the classical theory. The results show that for extremely fast reaction to occur in H_2/O_2 mixtures (nondilute or dilute), the temperature of the mixture has to exceed this transition temperature for any given pressure. This transition may occur during the induction time or during the consumption of major reactants.

The stability-sensitivity eigenanalysis provided a convenient means to identify this phenomena and more importantly quantify the differences between "weak" and "strong" ignition/reaction. Due to the fact that the system was driven by a single eigenvalue, the sensitivities of this eigenvalue and its associated eigenvector provided all the information necessary for understanding the controlling reactions of the mechanism. Although not a goal of this article, the eigenanalysis of the Green's function matrix produced overall reaction vectors which may be used to gain insight into mechanism reduction and lumping. Eigenanalyses of other matrices have been used for this purpose previously [38,39].

Acknowledgment

The authors acknowledge support from the Air Force Office of Scientific Research and the Office of Naval Research.

Appendix A. Degenerate Sensitivity Analysis

Equations (21) and (22) give the eigenvalue and eigenvector sensitivities provided that the system is nondegenerate. Consider now the degenerate case in which the matrix \underline{G} has a portion of its eigenvalues which are degenerate

$$(A-1(a)) \quad \underline{G}\underline{U}_i = E\underline{U}_i \quad i = 1, \dots, S$$

$$(A-1(b)) \quad \underline{G}\underline{U}_i = \lambda_i \underline{U}_i \quad i = S + 1, \dots, N$$

where E is the eigenvalue of S fold degeneracy and the remaining eigenvalues $\lambda_{S+1}, \lambda_{S+2}, \dots, \lambda_N$ are assumed to be nondegenerate. Without loss of generality, the first S eigenvectors are chosen as the degenerate set. The

eigenvectors \underline{U}_i in eq. (A-1(a)) are a particular, but nonunique set. Indeed, there is an infinite number of eigenvectors which would satisfy eq. (A-1(a)) by simply taking linear combinations of this particular set. This ambiguity causes difficulty when calculating the sensitivities of eigenvalues and eigenvectors.

Consider now the degenerate portion of the eigenvectors in eq. (A-1(a)) and again make exactly the same expansion as implied in eqs. (15), (16), and (17) except in this case

$$(A-2) \quad \underline{G} \longrightarrow \underline{G}(\underline{\alpha}) + \frac{d\underline{G}(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha}$$

$$(A-3) \quad E_\ell \longrightarrow E(\underline{\alpha}) + \frac{dE_\ell(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha} \quad \ell = 1, \dots, S$$

$$(A-4) \quad \underline{\phi}_\ell \longrightarrow \underline{\phi}_\ell(\underline{\alpha}) + \frac{d\underline{\phi}_\ell(\underline{\alpha})}{d\underline{\alpha}} \cdot d\underline{\alpha}$$

where

$$(A-5) \quad \underline{\phi}_\ell = \sum_{m=1}^S a_{\ell m} \underline{U}_m$$

is as yet an arbitrary linear combination of the degenerate eigenvectors. This produces a set of equations analogous to eq. (19) with the following form

$$(A-6(a)) \quad \underline{G} \underline{\phi}_\ell = E \underline{\phi}_\ell$$

$$(A-6(b)) \quad [\underline{G} - \underline{1}E] \frac{d\underline{\phi}_\ell}{d\underline{\alpha}} \cdot d\underline{\alpha} = \left[\underline{1} \frac{dE_\ell}{d\underline{\alpha}} \cdot d\underline{\alpha} - \frac{d\underline{G}}{d\underline{\alpha}} \cdot d\underline{\alpha} \right] \underline{\phi}_\ell$$

for $\ell = 1, 2, \dots, S$. In eq. (A-6(b)), an ambiguity exists because of the arbitrariness in the degenerate set of eigenvectors $\underline{\phi}_\ell$. A unique specification of the eigenvalues can only be achieved by giving a specific perturbation $(d\underline{G}/d\underline{\alpha}) \cdot d\underline{\alpha}$ since different perturbations would correspond to different possible zeroth order unperturbed degenerate eigenvectors $\underline{\phi}_\ell$. Therefore, the differential variation $d\underline{\alpha}$ cannot be removed from eq. (A-6(b)). Multiplication of eq. (A-6(b)) on the left by ${}_i \underline{\phi}^{-1}$, $i' = 1, 2, \dots, S$ with ${}_i \underline{\phi}^{-1} \cdot \underline{\phi}_i = \delta_{ii'}$ shows that the perturbed eigenvalues may be chosen to diagonalize the perturbation matrix

$$(A-7) \quad \frac{dE_i}{d\underline{\alpha}} \cdot d\underline{\alpha} \delta_{ii'} = {}_i \underline{\phi}^{-1} \cdot \left[\frac{d\underline{G}}{d\underline{\alpha}} \cdot d\underline{\alpha} \right] \cdot \underline{\phi}_i, \quad i = 1, \dots, S$$

This equation is the degenerate analog of eq. (21). Solution for the perturbed eigenvalues $(dE_i/d\underline{\alpha}) \cdot d\underline{\alpha}$ from eq. (A-7) will also yield a particular linear combination $\underline{\phi}_\ell$ of degenerate eigenvectors in eq. (A-5). In a similar fashion, multiplying eq. (A-6(b)) on the left by ${}_i \underline{U}^{-1}$, $i' = S + 1, \dots, N$ yields

$$(A-8) \quad \frac{d\underline{\phi}_i}{d\underline{\alpha}} \cdot d\underline{\alpha} = - \sum_{i'=S+1}^N \underline{U}_{i'} \left[{}_{i'} \underline{U}^{-1} \cdot \left[\frac{d\underline{G}}{d\underline{\alpha}} \cdot d\underline{\alpha} \right] \cdot \underline{\phi}_i \right] / [\lambda_{i'} - E] \quad i = 1, \dots, S$$

Equation (A-8) is the degenerate analog of eq. (22).

At this point, several comments need to be made. First, in eq. (22) the summation i' covers all of the degenerate and nondegenerate states except as indicated in the summation. However, when the sum runs over the degenerate states it is necessary to include the following replacement

$$\lambda_{i'} \longrightarrow E \quad \text{and} \quad \underline{U}_{i'} \longrightarrow \underline{\phi}_{i'}.$$

The latter replacement just insures that the proper superposition of the degenerate states is utilized. The derivatives of the eigenvalues and eigenvectors in eqs. (A-7) and (A-8), respectively, for the degenerate case are sometimes referred to as directional derivatives since they require a particular specification of a differential parameter change $d\alpha$. Assuming the parameters individually have a distinct physical meaning, the natural choice is to perform the analysis sequentially with the separate choices $d\alpha \equiv d\alpha_1$, $d\alpha \equiv d\alpha_2, \dots$ etc. Note that in the latter case of a single parameter change, the differential term in eq. (A-6(b)) may again be removed but it always must be understood that the resultant sensitivities correspond to that particular differential parameter change. Note also that the restrictions on the summations in eqs. (A-8) and (22) remove what would otherwise be another ambiguity in the eigenvector derivations. In particular, these summation restrictions specify that the eigenvector derivatives have no components along the corresponding unperturbed ones and this is sometimes referred to as a specification of normalization.

Bibliography

- [1] S.G. Saytzev and R.I. Soloukhin, *Eighth Symposium (International) on Combustion*, Williams and Wilkins, Eds., 1962, p. 344; R. A. Strehlow and A. Cohen, *Phys. Fluids*, **5**, 97 (1962).
- [2] V.K. Baev, V.I. Golovichev, V.I. Dimitrov, R.I. Soloukhin, and V.A. Yasakov, *Fizika Goreniyaa i Vzryva*, **9**, 823 (1973).
- [3] E.S. Oran, T.R. Young, J.P. Boris, and A. Cohen, *Combustion and Flame*, **48**, 135 (1982).
- [4] W.C. Gardiner, Jr. and C.B. Wakefield, *Astronautica Acta*, **15**, 399 (1970).
- [5] V.V. Voevodsky and R.I. Soloukhin, *Eighth Symposium (International) on Combustion*, Williams and Wilkins, Baltimore, 1962, p. 335.
- [6] J.W. Meyer and A.K. Oppenheim, *Thirteenth Symposium (International) on Combustion*, The Combustion Institute, Pittsburgh, Pennsylvania, 1970, p. 279.
- [7] E.S. Oran and J.P. Boris, *Combustion and Flame*, **48**, 149 (1982).
- [8] R.M. Hedges and H. Rabitz, *J. Chem. Phys.*, **82**, 3674 (1985).
- [9] R.A. Yetter, F.L. Dryer, and H. Rabitz, *Combust. Sci. Tech.*, in press, 1990.
- [10] D.K. Stull and H. Prophet, Eds., *JANAF Thermochemical Tables*, NSRDS-NBS 37, 1971; also Dow Chemical Co., Midland, Michigan, distributed by Clearing House for Federal Scientific and Technical Information, PB 168370, 1965. Also see M.W. Chase, Jr., C.A. Davies, J.R. Downey, Jr., D.J. Fulrip, R.A. McDonald, and A.N. Syverud, *JANAF Thermochemical Tables, Third Edition*, *J. Phys. Chem. Ref. Data*, **14**, Supplement 1 (1985).
- [11] L.G.S. Shum and S.W. Benson, *J. Phys. Chem.*, **87**, 3479 (1983).
- [12] S. Gordon and B.J. McBride, *NASA SP-273, Interim Revision*, 1976.
- [13] R.J. Kee, F.M. Ripley, and J.A. Miller, *Sandia Report SAND87-8215*, Livermore, California, 1987.
- [14] E.R. Ritter and J.W. Bozzelli, Dept. of Chemical Engineering, Chemistry, and Environmental Science, New Jersey Institute of Technology, Newark, New Jersey, March 16, 1987.
- [15] A.J. Hills and C.J. Howard, *J. Chem. Phys.*, **81**, 4458 (1984).
- [16] P.D. Lightfoot, B. Veyret, and R. Lesclaux, *Chem. Phys. Letters*, **1**, 120 (1988).
- [17] L. Brouwer, C.J. Cobos, J. Troe, H.-R. Duba, and F.F. Crim, *J. Chem. Phys.*, **86**, 6171 (1987).

- [18] A. N. Pirraglia, J.V. Michael, J.W. Sutherland, and R. B. Klemm, *J. Phys. Chem.*, **93**, 282 (1989).
- [19] J.W. Sutherland, J.V. Michael, A. N. Pirraglia, F.L. Nesbitt, and R. B. Klemm, *Twenty-first Symposium (International) on Combustion*, The Combustion Institute, Pittsburgh, Pennsylvania, 1986, p. 929.
- [20] J.V. Michael and J.W. Sutherland, *J. Phys. Chem.*, **92**, 3853 (1988).
- [21] W. Tsang and R. F. Hampson, *J. Phys. Chem. Ref. Data*, **15**, 1987 (1986).
- [22] M.W. Slack, *Combustion and Flame*, **28**, 241 (1977).
- [23] J. Warnatz, in *Combustion Chemistry*, W.C. Gardiner, Jr., Ed., Springer-Verlag, New York, 1985.
- [24] A.C. Hindmarsh, *ACM SIGNUM Newsletter*, **15**, 10 (1980).
- [25] R. J. Kee, J. A. Miller, and T. H. Jefferson, *Sandia Report SAND80-8003*, Sandia National Laboratories, Livermore, California, 1980.
- [26] G. B. Skinner and G. H. Ringrose, *J. Chem. Phys.*, **42**, 2190 (1965).
- [27] G. L. Schott and J. L. Kinsey, *J. Chem. Phys.*, **29**, 1177 (1958).
- [28] R. A. Yetter, F.L. Dryer, and H. Rabitz, *Combust. Sci. Tech.*, in press.
- [29] I. N. Levine, *Quantum Chemistry*, 2nd Ed., Allyn and Bacon, Boston, 1974, p. 371.
- [30] M. Mishra, L. Peiperl, Y. Reuven, H. Rabitz, R. A. Yetter, and M. D. Smooke, *J. Chem. Phys.*, in press.
- [31] M. A. Kramer, J. M. Calo, H. Rabitz, and R. J. Kee, *Sandia Report SAND82-8231*, Sandia National Laboratories, Livermore, California, 94550, 1982.
- [32] B. Lewis and G. vonElbe, *Combustion, Flames, and Explosion of Gases*, 2nd Ed., Academic Press, New York, 1961.
- [33] R. R. Baldwin, D. Jackson, R.W. Walker, and S. J. Webster, *Trans. Farad. Soc.*, **63**, 1665 (1967).
- [34] R. R. Baldwin, D. Jackson, R.W. Walker, and S. J. Webster, *Trans. Farad. Soc.*, **63**, 1676 (1967).
- [35] G. Dixon-Lewis and D. J. Williams, *Comprehensive Chemical Kinetics*, C. H. Bamford and C. F. H. Tipper, Eds., Elsevier, Amsterdam, 1977, p. 1-248.
- [36] E. P. Dougherty and H. Rabitz, *J. Chem. Phys.*, **72**, 6571 (1980).
- [37] U. Maas and J. Warnatz, *Combustion and Flame*, **74**, 53 (1988).
- [38] S. H. Lam and D. A. Goussis, *Twenty-second Symposium (International) on Combustion*, The Combustion Institute, Pittsburgh, 1988, p. 931.
- [39] S. Vajda, H. Rabitz, and R. A. Yetter, *Combustion and Flame*, **82**, 270 (1990).

Received June 11, 1990

Accepted October 19, 1990

Appendix D

4. On the Use of Green's Functions for the Analysis of Dynamic Couplings: Some Examples of Chemical Kinetics and Quantum Dynamics, M. Mishra, L. Peiperl, Y. Reuven, H. Rabitz, R. Yetter, and M. Smooke, J. Phys. Chem., 95, 3109 (1991).

On the Use of Green's Functions for the Analysis of
Dynamic Couplings: Some Examples from
Chemical Kinetics and Quantum Dynamics

Manoj Mishra, Lawrence Peiperl,
Yakir Reuven and Herschel Rabitz
Department of Chemistry
Princeton University
Princeton, NJ 08544

and

Richard A. Yetter
Department of Mechanical and
Aerospace Engineering
Princeton University
Princeton, NJ 08544

and

Mitchell D. Smooke
Department of Mechanical Engineering
Yale University
New Haven, CT 06520

ABSTRACT

The utility of individual elements of Green's function matrices, in the investigation of dynamic couplings, is illustrated by offering examples from linear and nonlinear kinetics and quantum dynamics. The concept of reduced Green's functions affords a detailed characterization of the actual pathways mediating these couplings. Self similar behavior between different elements of the Green's function matrix indicates the presence of strong coupling between different variables of the model. We investigate the structure of the entire Green's function matrix to examine such self similar behavior and other simplifying characteristics of concern for physical insight as well as for economic modeling of the dynamic systems. Global structure in the entire Green's function matrix may be used to reduce the complexity (number of dependent variables) in a model.

I. Introduction

Green's functions are traditionally used as a means for solving linear models driven by inhomogeneous source terms. The interpretation of Green's functions as response functions underlies their use in propagator based methods of Quantum Mechanics.¹ While the residues and poles of the Green's functions have found extensive use in spectral analyses,² the use of Green's functions for investigating the coupling between different variables of dynamical systems has found limited applications so far.³ In this paper, we offer examples of their use in a diverse set of complex chemical/physical problems to call attention to the power and efficacy of these functions in deciphering the latent dynamic couplings, generally masked by the complex network structure in the model.

Section II.a will first examine the role of Green's functions as response functions by identifying them as sensitivity coefficients of the model. The new concept of reduced Green's functions affords a detailed characterization of the complex dynamics and is discussed in Section II.b. Section III presents illustrative examples of Green's functions and some related reduced Green's functions from nonlinear kinetics problems, including as well as excluding transport, and emphasizes their use in revealing latent system couplings. Further examples from some model problems in quantum dynamics and linear kinetics are presented in Section IV. The diverse examples underscore the universal utility of these concepts. In dynamical systems with strong coupling, dominant control of a dependent variable can result in self similar behavior between the different elements of the Green's function matrix. Examples from the use of the entire Green's function matrix for seeking simplifying features of the complex network of elementary steps in kinetics and their use in formulating more

tractable models are offered in Section V. A brief summary of our findings concludes the paper.

II.a Green's Functions as Response Functions

To best understand Green's functions from diverse chemical problems we consider cases where the physical phenomena are described by a vector set of differential equations

$$L(Q, \underline{a}) = 0 \quad (II.1)$$

Here Q is the sought after vector of dependent variables (e.g., concentration profiles in kinetics, amplitudes in quantum mechanics or the canonically conjugate variables of classical Hamiltonian dynamics) and L_i is an element of the appropriate differential operator vector for the respective problem. The elements of the vector \underline{a} constitute the system's physical parameters (e.g., rate constants and diffusion coefficients in kinetics, potential surface parameters in dynamics, etc.). The spatial and/or temporal dependence of the solution vectors is not explicitly shown for clarity and is assumed to be known numerically through the solution of the system of equations (II.1), augmented by appropriate initial and/or boundary conditions. Sections III and IV will provide specific physical illustrations of Eq. (II.1).

To establish the physical content of the system Green's function we modify Eq. (II.1) by the addition of an incremental flux term δJ_i at time t and position x (we shall just consider one dimensional spatial problems for simplicity of illustration) as a source for the i^{th} equation.

$$L_i(Q, \underline{a}) = \delta J_i(x, t) \quad (II.2)$$

Functional differentiation of Eq. (II.2) with respect to the new added flux terms leads to

$$\sum_n \left(\frac{\partial L_i}{\partial O_n} \right) \left(\frac{\delta O_n}{\delta J_{n'}} \right) = \delta_{in'} \delta(x-x') \delta(t-t') \quad (II.3)$$

Here the Green's function matrix elements $G_{nn'}(x,t;x',t') = \delta O_n(x,t)/\delta J_{n'}(x',t')$ are functional derivatives and provide the response of the n^{th} dependent variable at (x,t) to a change in the flux of the n'^{th} dependent variable at a prior time t' and position x' . This statement is explicitly evident from the first order functional Taylor expansion implied by Eqs. (II.2) and (II.3) to produce⁴

$$\delta O_n(x,t) = \sum_{n'} \int dx' \int dt' G_{nn'}(x,t;x',t') \delta J_{n'}(x',t') \quad (II.4)$$

$$\frac{\delta O_n}{\delta \alpha_i} = \sum_{n'} \int dx' \int dt' G_{nn'}(x,t;x',t') \partial L_{n'}(x',t') / \partial \alpha_i \quad (II.5)$$

The identification of the solution to Eq. (II.3) as a Green's function may be made regardless of whether Eq. (II.1) is a linear equation. A Green's function is associated with the linear differential equations driven by the Jacobian, $\partial L_i / \partial O_n$, in Eq. (II.3). A basic application of the system Green's function is to provide a closed form expression for the parametric sensitivity coefficients, although this latter application is not the focus of the present paper.

In the case of pure temporal kinetics, allowing for discrete parametric variations only, the identity of $G_{nn'}(t,t') = \delta O_n(t) / \delta J_{n'}(t')$ is

easily established. As a convenient shorthand notation, the pure temporal Green's function $G_{nn'}(t, t')$ is sometimes written as $\partial O_n(t)/\partial O_{n'}(t')$. In a similar fashion, a steady state Green's function may also be identified as having the elements $G_{nn'}(x, x') = \delta O_n(x)/\delta J_{n'}(x')$ with a similar interpretation.

In the case of Heisenberg's equation of motion for the time evolution operator, the Green's function $G(t, t')$ for the corresponding sensitivity equations is well known to be the time evolution operator itself.⁵ The i, j matrix element of the time evolution operator represents the transition amplitude between eigenstates i and j as driven by the coupling in the Hamiltonian. These features are discussed in detail in a following section.

From Eq. (II.4) it is evident that the Green's function matrix determines the stability of a dynamical system: a large magnitude of G_{ij} being indicative of instability with respect to changes in the flux of the j^{th} dependent variable. In the case of pure temporal systems, since for reasons of causality the disturbance $\delta J_j(t')$ must precede the response $\delta O_i(t)$, the relation of Green's functions to stability analysis and control theory becomes readily apparent.⁶ (Analogous arguments also apply to the temporal dependence of space-time systems). The eigenvalues of the \underline{G} matrix (actually their logarithms) may be identified as time-dependent Lyapunov exponents⁷

$$\lambda_n(t, t') = \underline{U}_n^T(t, t') \underline{G}(t, t') \underline{U}_n(t, t') \quad (\text{II.6})$$

where $\underline{U}_n(t, t')$ is the n^{th} eigenvector and $\lambda_n(t, t')$ is the associated eigenvalue of \underline{G} . Dynamic instability is indicated by any of the eigenvalues satisfying $|\lambda_n| > 1$. These latter quantities depend on the

current time as well as the time of the initial condition specification, thus indicating a retention of system integrated time history. One may also probe for which physical variables contribute to the system stability⁶ by differentiating Eq. (II.6) with respect to a system parameter to produce $\partial\lambda_n(t,t')/\partial\alpha_j$. An accompanying expression for the eigenvector sensitivities may also be established. The critical nature of this information is specially important when parameters are of a design nature and controllable in the laboratory.

In the case of the steady state Green's functions⁸ $G_{ij}(x,x')$, the presence of any eigenvalue satisfying $|\lambda_n|>1$ would imply that the dynamic system is not at a stable steady-state. In such a case, the full spatio-temporal problem should be solved and propagated sufficiently far in time to achieve a stable steady-state solution.

In any problem where the dependent variables are directly measurable or controllable, then the Green's function elements themselves may also be measured. This measurement, for example in kinetics, could be achieved by disturbing a given species (or eigenstate in quantum dynamics) and monitoring the response amongst all of the other species (or eigenstates). In this way, it may be possible to determine how to alter the spatial or temporal response of a system by a judicious use of Green's functions.

II.b Reduced Green's Functions

While the elements of the Green's function matrix provide information about the coupling between the dependent variables, they do not reveal the pathway of coupling. As a concrete example, consider the case of pure temporal kinetics governed by the equation

$$\frac{dO_i}{dt} = f_i(Q, \underline{Q}) \quad (II.7)$$

The right-hand side of Eq. (II.7) contains all of the information about the kinematic coupling in the system, but the actual dynamic coupling may differ due to complex nonlinear interactions only present in the solution to the equation. The magnitude of G_{ij} indicates if O_i and O_j are coupled but it does not tell us whether a third (or several other) dependent variables mediate the response. In other words, the pathway or dynamic coupling is not evident from examining the original differential equations, nor is it revealed by the fundamental Green's function \underline{G} alone.

This detailed pathway insight into the actual modes of coupling is provided by an analysis of the reduced Green's functions. Such an analysis is carried out by considering variations of only a portion of the dependent variables while holding another portion constrained as fixed. Therefore, upon consideration of the dependent variable vector, we may partition it into two parts $\underline{Q} = (\underline{Q}', \underline{Q}'')$ where variations of the second portion are constrained to be $\delta \underline{Q}'' = \underline{0}$. Accordingly, we may calculate the elements of the reduced Green's function

$$G'_{ij} = \left. \frac{\delta O_i}{\delta J_j} \right|_{\delta \underline{Q}'' = \underline{0}} \quad (II.8)$$

where this constrained matrix satisfies an equation of exactly the same form as Eq. (II.3), except that now the Jacobian is of reduced dimension with the columns and rows associated with \underline{Q}'' removed. Elements of this reduced matrix probe the system's dynamic response where all couplings mediated by \underline{Q}'' have been disabled. It should be emphasized that while \underline{Q}''

have been frozen, they have not been deleted from the problem and their nominal values obtained from the solution of Eq. (II.1) are retained in the reduced calculation. Only their response to variations of \underline{Q}' is not allowed. A judicious partitioning of \underline{Q} into \underline{Q}' and \underline{Q}'' , followed by an examination of the corresponding reduced Green's function, is a useful tool for deciphering the dynamic couplings responsible for the system behavior.

In the following sections, we offer examples from several problems to illustrate these varied roles of the Green's function and some related, reduced Green's functions.

III. Green's Functions for Pure Temporal Reactions and Reaction-Convection-Diffusion Systems

The general class of problems treated in this category may be described by the following reaction-diffusion-convection equation

$$\dot{m} = \rho u = \text{constant}$$

$$\rho \frac{\partial O_k}{\partial t} = \frac{\partial}{\partial x} \left(\rho D_k \frac{\partial O_k}{\partial x} \right) - \dot{m} \frac{\partial O_k}{\partial x} + f_k(Q, \underline{\alpha}, T) \quad (\text{III.1a})$$

$$\rho \frac{\partial T}{\partial t} = \frac{1}{C_p} \frac{\partial}{\partial x} \left(\lambda \frac{\partial T}{\partial x} \right) - \dot{m} \frac{\partial T}{\partial x} + \frac{1}{C_p} \sum_{k=1}^N \rho D_k C_{pk} \frac{\partial O_k}{\partial x} \frac{\partial T}{\partial x} + \frac{H(Q, \underline{\alpha}, T)}{C_p} \quad (\text{III.1b})$$

$$0 \leq x \leq L \quad (\text{III.1c})$$

where for simplicity we confine ourselves to considering only one spatial dimension. In this equation, O_k is the mass fraction of the k^{th} species, T is the temperature, \dot{m} is the mass flow rate, D_k is the diffusion coefficient of the k^{th} species with respect to the mixture, f_k is the rate of production/destruction of the k^{th} species, H is the reactive enthalpy term, λ is the mixture thermal conductivity, C_p is the constant pressure heat capacity with individual components C_{pk} , u is the velocity and ρ is the mass density of the mixture. The vector $\underline{\alpha}$ represents the remaining system parameters (e.g., activation energies, Arrhenius pre-exponential factors, etc.). The system of Eqs. (III.1) is supplemented by requisite initial and boundary conditions, an equation of state where appropriate an equation for the conservation of momentum may also be prescribed.

Reaction-convection-diffusion models defined by Eq. (III.1) involving both temporal evolution and spatial transport are difficult to solve and two natural restricted cases A and B below have seen maximum activity:

- A. The pure temporal case without spatial diffusion or convection described by

$$\rho \frac{dO_k}{dt} = f_k(Q, \alpha, T) \quad (\text{III.2a})$$

$$\rho \frac{dT}{dt} = \frac{H(Q, \alpha, T)}{C_p} \quad (\text{III.2b})$$

along with a set of initial conditions

$$O_k(0) = O_k^0; \quad T(0) = T^0 \quad (\text{III.3})$$

- B. The steady state limit without any time dependence described by:

$$\dot{m} = \rho u = \text{constant} \quad (\text{III.4a})$$

$$\dot{m} \frac{dO_k}{dx} = \frac{d}{dx} \left(\rho D_k \frac{dO_k}{dx} \right) + f_k(Q, \alpha, T) \quad (\text{III.4b})$$

$$\dot{m} \frac{dT}{dx} = \frac{1}{C_p} \frac{d}{dx} \left(\lambda \frac{dT}{dx} \right) + \frac{1}{C_p} \sum_{k=1}^N \rho D_k C_{pk} \frac{\partial O_k}{\partial x} \frac{\partial T}{\partial x} + \frac{H(Q, \alpha, T)}{C_p} \quad (\text{III.4c})$$

$$0 \leq x \leq L$$

In the case of a premixed laminar flame the appropriate boundary conditions at $x=0$ are

$$T(0) = T_0, \quad O_k - \rho D_k \frac{\rho D_k}{\dot{m}} \frac{dO_k}{dx} = \epsilon_k \quad (\text{III.5a})$$

and at $x = L$

$$\left. \frac{dT}{dx} \right|_L = 0, \quad \left. \frac{dO_k}{dx} \right|_L = 0 \quad (\text{III.5b})$$

where T_0 is the temperature of the unreacted gas (for details, see ref. 9). If the problem is adiabatic, then λ is an eigenvalue and an additional boundary condition is needed.

The Green's function for these particular limiting situations satisfy special cases of (II.3). For Eqs. (III.2,3) we have

$$\left(\frac{\partial}{\partial t} - \underline{J} \right) \underline{G}(t, t') = \underline{1} \delta(t - t') \quad (\text{III.6})$$

where $J_{ij} = \partial f_i / \partial O_j$, $\underline{G}(t', t') = \underline{1}$ and for reasons of causality, $G_{ij}(t, t') = 0$ for $t < t'$. For the steady state limit described by Eqs. (III.4,5), the system Green's function is defined to satisfy the following equation

$$\left(\frac{\partial}{\partial x} \left(\rho \underline{D} \frac{\partial}{\partial x} \right) - \underline{1} \underline{m} \frac{\partial}{\partial x} + \underline{J} \right) \underline{G}(x, x') = \underline{1} \delta(x - x') \quad (\text{III.7})$$

and $\underline{G}(x, x') = \underline{0}$ for x, x' on the boundaries (0 or L) with \underline{D} being a diagonal matrix of diffusion coefficients.

Various strategies for the solution of Eqs. (III.6) and (III.7) have been reviewed elsewhere,³ and we will instead focus upon examples establishing the utility of Green's functions in investigating the dynamic couplings not readily discernible from a knowledge of the underlying kinematic mechanism alone.

As the first example, we consider the temporal kinetics of the wet oxidation of carbon monoxide. A comprehensive reaction mechanism¹⁰ for describing this process is given in Table I. An inspection of Table I reveals that several elementary steps directly participate in the consumption of carbon monoxide. At intermediate and high temperatures, it is well established that the major consumer of carbon monoxide is the hydroxyl

radical through reaction 11. At lower temperatures (below ~900 K), reaction 9, with the hydroperoxy radical, may dominate. The exact role of other intermediates of the system (e.g., H, O, H₂O₂, HCO, etc.) on the kinetics of the oxidation process is very difficult to discern from the mechanistic (kinematic) data of Table I since for some intermediates a direct consumption reaction does not exist, while for others, the rate constants are very small.

Consider, for example, the correlation of carbon monoxide with hydrogen atoms. The only elementary step involving the direct reaction of H and CO is reaction 52. However, the thermodynamically favored direction of this reaction is the reverse reaction 51. Indirectly, the H atom is involved in the production and consumption of the important hydroxyl and hydroperoxy radicals (e.g., through steps 15, 18, 48, etc.). Due to a variety of chain branchings in the reaction mechanism, the indirect effect of H upon CO could be quite significant. Brute force estimation of the coupling between H and CO would necessitate repeated solution of Eqs. (III.2) for a variety of H atom concentrations and at different initial times. The use of the nominal and reduced Green's functions obviates this laborious investigation and provides quantitative information about the desired couplings.

To illustrate this point, we present two Green's function response surfaces, $\delta\text{CO}(t)/\delta J_{\text{H}}(t')$ and $\delta\text{CO}(t)/J_{\text{OH}}(t')$, in Figure 1, for a dilute carbon monoxide-water-oxygen mixture reacting homogeneously and isothermally in nitrogen at 1100 K and 1 atmosphere. These results were obtained by solving Eqs. (III.2a and 6), using the stiff ODE numerical code of Hindmarsh¹¹ in combination with the Green's function code of Kramer, et al.¹² More details on the specific calculations may be found in Ref. 13.

Both Green's function surfaces exhibit pronounced negative response in the vicinity of $t \approx 10^{-2}$ sec. This latter time corresponds closely to the maximum in the radical pool concentration profiles. Interestingly, the coupling of the CO concentration with the H-atom concentration is $\sim 50\%$ greater than the coupling with OH, despite the fact that the latter species is the primary oxidant. Moreover, both response surfaces, as a function of time, are essentially identical in shape and therefore the physical implications from disturbing either the H concentration or the OH concentration are the same. For example, the response surface of $\delta CO(t)/\delta J_H(t')$ implies that if the H-atom is perturbed at or after $t' \sim 10^{-2}$ sec, no significant changes are predicted in the CO concentration at any time. For perturbations prior to $t' \sim 10^{-2}$ sec, the CO concentration first exhibits no response during the induction period, then rapidly achieves a negative peak and decays to zero. It is clear that late in the reaction, the CO concentration displays a "loss of memory" to early H or OH perturbations. Even perturbations in the H_2O_2 and HCO concentrations which do not directly consume carbon monoxide have similar response surfaces to those in Figure 1 with the magnitude of responses nearly the same as $\delta CO(t)/\delta J_{OH}(t')$. This type of self-similar behavior is a result of strong coupling amongst the members of the radical pool and this issue will be discussed further in section V.

A more detailed investigation of the coupling pathways can be obtained by calculating the reduced Green's functions, for example, with OH constrained to its nominal profile. In Figure 2 we present the $t'=0$ cuts of reduced Green's functions for the response of CO from which it is clear that the strong dynamic coupling between CO and H (Figure 1) is eliminated by freezing the OH profile at its nominal value (i.e., the strong response

of Figure 1 is now reduced to a weak broad profile). It is therefore clear that the carbon monoxide-hydrogen coupling results indirectly through the hydroxyl radical. Furthermore, it becomes apparent that the direct coupling by recombination reaction 51 plays a relatively insignificant role. The importance of the OH radical is further underscored in Figure 2 by the drastically reduced magnitudes of the maxima of responses to perturbations in the flux HO_2 , H_2O_2 and O, in comparison to their corresponding unconstrained curves (not shown here).

A similar illustration can also be given for the analogous reaction-convection-diffusion problem. These calculations correspond to a laminar premixed $\text{CO}/\text{H}_2/\text{O}_2$ flame. Details of the calculations are presented elsewhere.¹⁴ The calculations are based on the same reaction mechanism of Table I using the numerical code of reference 9. The Green's function coefficients for $\delta\text{CO}(x)/\delta J_{\text{H}}(x')$ and $\delta\text{O}(x)/\delta J_{\text{OH}}(x')$ are shown in Figure 3. Here, the maximum response of CO to the perturbation of H-atom flux is approximately 20 times larger than that due to the perturbation of the OH flux. The flux perturbation in H and OH concentrations occurs along the diagonal $x=x'$ and consequently any variation in the CO concentration at position $x' > x$ exists due to upstream transport by diffusion with simultaneous chemical reaction. The maximum response of the CO in position x occurs in the immediate vicinity of the flame front with a broad secondary response both upstream and downstream in the flow. The magnitude of the results of Figure 3 are consistent with the fact that H-atoms diffuse more readily than OH radicals and strongly suggest that the role of transport in the $\text{CO}+\text{H}_2+\text{O}_2$ chemistry may be much more important than believed so far.¹⁴ The self-similar behavior of the OH and H response surfaces once again indicates strong coupling between different variables and is easily

understood in terms of the scaling and self similarity relations to be discussed in Section V.

The freezing of the OH response again significantly affects the coupling between the CO concentration and the H-atom concentration (Figure 4). Here the reduced Green's function $\delta U(x)/\delta J_H(x')|_{\delta OH=0}$ shows that the introduction of a small flux of H-atoms will inhibit the CO consumption, whereas in the temporal problem, the overall reaction was still accelerated but by a significantly reduced amount.

IV. Green's Functions from Quantum Dynamics and Linear Kinetics

In the examples cited in the previous section, the fundamental and reduced Green's functions were found to be valuable in the analysis of intricate couplings resulting from the nonlinearity of the governing Eqs. (III.2,4). Their use can identify the extent of coupling between various species, as well as any mediators of these couplings. The information so obtained can run counter to the expectations from the reaction network structure alone. While the unforeseeable nature of the dynamic couplings in chemical kinetics may be attributed to the nonlinear nature of the mass action kinetics, even linear governing equations, such as in quantum mechanics, can lead to dynamic couplings which cannot be anticipated by knowledge of the Hamiltonian coupling alone. It is therefore useful to explore the utility of the fundamental and reduced Green's functions in the analysis of dynamic couplings in quantum phenomena as well as linear kinetics.

A. *Quantum Mechanics*

We can study quantum dynamics as an evolution of probability amplitude or equivalently under the influence of some perturbation V acting amongst the zeroth order eigenstates of a time independent Hamiltonian H_0 . The nature and extent of this amplitude flow is determined by the time evolution matrix $\underline{U}(t,t')$, which is governed by the following equation of motion¹⁵

$$\frac{d}{dt}\underline{U}(t,t') = \frac{-i}{\hbar}H(t)\underline{U}(t,t') \quad (\text{IV.1})$$

where $H(t) = H_0 + V(t)$ is the time dependent Hamiltonian. The initial condition for Eq. (IV.1) is

$$\underline{U}(t,t') = \underline{1} \quad (\text{IV.2})$$

and we have assumed that the eigenbasis for H_0 is used for representing the operators. The Green's function $\underline{G}(t, \tau; \underline{\alpha})$ satisfies

$$\left[\underline{1} \frac{d}{dt} + \frac{i}{\hbar} \underline{H}(t) \right] \underline{G}(t, \tau) = \underline{1} \delta(t - \tau) \quad (\text{IV.3})$$

$$\underline{G}(\tau, \tau) = \underline{1} \quad (\text{IV.4})$$

A comparison of Eqs. (IV.3) and (IV.1) and their initial conditions shows that the Green's function $\underline{G}(t, \tau)$ for $t > \tau$ is simply the time evolution operator $\underline{U}(t, \tau)$. A nonvanishing $G_{ij} = U_{ij}$ implies dynamic coupling between eigenstates i and j , and once again, we see the role of the Green's function in reflecting dynamic couplings. Although the time evolution operator is well known in quantum mechanics, its interpretation, in the sense discussed in this paper, is unusual, particularly in purely temporal analogues of Eqs. (II.4) and (II.5). Again Eq. (II.4), in this case, is simply a statement of the Green's function acting as a propagator for the evolution of an amplitude disturbance, while Eq. (II.5) shows that the Green's function dictates the temporal behavior of any parameter disturbance in the system Hamiltonian.

A quantity of general interest is the probability that application of the perturbation V at some time t' will lead to transition from eigenstate i to the eigenstate j of H_0 , where the measurements are done after an infinitely long period (compared to the time scale of the internal motions of the system). We therefore focus our interest on the long time average

$$\langle |G_{ij}|^2 \rangle_{t \rightarrow \infty} = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t |G_{ij}(t, t')|^2 dt' \quad (\text{IV.5})$$

When the system is described by a time independent Hamiltonian, as in the examples below, the Green's function is given by

$$\underline{G}(t, t'; \underline{\alpha}) = \exp \left[-\frac{i}{\hbar} (t - t') \underline{H}(\underline{\alpha}) \right] \quad (\text{IV.6})$$

In terms of the eigenvectors \underline{T} and the diagonal matrix of eigenvalues \underline{h} of \underline{H} , we have

$$\underline{H} \underline{T} = \underline{T} \underline{h} \quad (\text{IV.7})$$

$$\underline{G}(t, t') = \underline{T} \exp \left[-\frac{i}{\hbar} \underline{h} (t - t') \right] \underline{T} \quad (\text{IV.8})$$

$$G_{ij}(t, t') = \sum_k T_{ik} \exp \left[-\frac{i}{\hbar} h_{kk} (t - t') \right] T_{jk} \quad (\text{IV.9})$$

and the long time average becomes

$$\begin{aligned} \langle |G_{ij}|^2 \rangle_{\tau \rightarrow \infty} &= \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_m \sum_k T_{ik} T_{im} T_{jk} T_{jm} \int_{t'}^t dt \exp \left[-\frac{i}{\hbar} (t - t') (h_{kk} - h_{mm}) \right] \\ &= \sum_k T_{ik}^2 T_{jk}^2 + \sum_m \sum_{k \neq m} T_{ik} T_{im} T_{jk} T_{jm} \\ &\quad h_{kk} - h_{mm} \end{aligned} \quad (\text{IV.10})$$

The computation of this average thus requires only the eigenvectors \underline{T} and eigenvalues \underline{h} of \underline{H} . We shall refer to the average $\langle |G_{ij}|^2 \rangle_{\tau \rightarrow \infty}$ as the mean square Green's function or average transition probability. The limitations that stem from the choice of an arbitrary basis to investigate dynamic coupling of states are well known.¹⁶ This limitation does not, however,

vitiates the qualitative insight gained from examining the structure of the Green's function especially when H_0 and V play physically distinct roles.

The kinematic coupling is revealed by the structure of the Hamiltonian matrix. As an example, we model the Hamiltonian of two coupled oscillators in Figure 5 with eight and two accessible eigenstates, respectively. In terms of a direct product (8×2) of eigenbases of the uncoupled oscillators, we have a 16 dimensional representation made up of two blocks corresponding to the two high frequency modes of one oscillator being coupled to the eight eigenstates of the other. The state $|8,1\rangle$ in which the first oscillator is in its highest frequency mode and the second in the lower of its two modes is directly coupled to the state $|1,2\rangle$ in which the first oscillator is in this lowest frequency mode and the second in the higher of its two modes. The diagonal elements increase as integers to mimic the energy levels of harmonic oscillators.

The long term dynamic coupling, as discerned from an examination of $\langle |G|^2 \rangle_{\tau \rightarrow \infty}$ is portrayed by Fig. 6. The simple kinematic coupling structure of Fig. 5 leads to dynamic couplings clearly unpredictable from the knowledge of the kinematic couplings alone.

Figure 7 represents a variation on the previous example in which both of the oscillators now have four eigenstates. The sharply banded structure associated with the mean square Green's function in Fig. 8 shows that, while coupling is by no means limited to directly connected states by the Hamiltonian, neither is dynamic coupling distributed equally among all of the states. This surprising structure reinforces the important role of Green's functions in the analysis of dynamic couplings.

In Figs. 9-12 we elucidate the use of the reduced Green's function method using a pentadiagonal Hamiltonian with elements ranging over three

orders of magnitude. The full Hamiltonian and the corresponding Green's function are presented first in Figs. 9 and 10, respectively. The reduced Green's functions for the same Hamiltonian in which the seventh state has been eliminated is shown in Fig. 11 and that in which the ninth state has been eliminated is portrayed by Fig. 12. Figure 11 shows that state 7 is a critical pathway for coupling between states 1-6 and 8-16, since its elimination virtually uncouples the two blocks. In contrast, Fig. 12 reveals that state 9 contributes only in a minor fashion to the overall dynamic coupling.

B. Linear Kinetics

The time evolution of species concentrations in linear temporal kinetics is described by

$$\frac{dQ}{dt} = \underline{M}Q, \quad Q(t_0) = Q^0 \quad (\text{IV.11})$$

which is analogous to the governing Eqs. (IV.1) of quantum dynamics except for the absence of $-i/\hbar$. The presence of $-i/\hbar$ leads to a rich interference between the probability amplitudes or Green's function elements for different eigenstates during the evolution of quantum mechanical systems. It is therefore useful to contrast the Green's functions from quantum dynamics and linear kinetics described by the same Jacobian ($\underline{M} = \underline{H}$).

Conservation of matter implies that all the off-diagonal elements of \underline{M} be positive and the elements of any column of \underline{M} must add up to zero. In addition, the matrices used here are real symmetric (and hence Hermitian) to double as an acceptable Hamiltonian. Physically this latter symmetry represents reactions at temperatures high enough to make the differences in forward and reverse activation barriers insignificant. Due to the absence

of $-i\hbar$ in the linear kinetics problem, the corresponding Green's function does not lend itself to the time averaging used in the case of quantum dynamical systems. We have instead studied them as a function of the time interval $(t-t')$.

In Fig. 13, we present the matrix which doubles as both $\underline{\underline{M}}$ and $\underline{\underline{H}}$. In linear kinetics, this matrix represents a cyclic reaction network where each species is directly coupled to the next, and the last is coupled back to the first. It is found that the behavior of the Green's function matrix elements G_{ij} depend only on the mode of coupling between the corresponding species. Since no two species are separated by more than five intermediaries, only six different types of plots are seen (Fig. 14). The entire Green's function matrix is represented in Table II to underline the interrelationships. It is seen that G_{ij} , for smaller values of $|i-j|$, have a much larger magnitude for times $(t-t') \leq 12$. On the other hand, the quantum mechanical mean square Green's function driven by the same matrix $\underline{\underline{H}} = \underline{\underline{M}}$ (Fig. 15) reveals that the interference structure leads to essentially uniform long-range coupling between all the eigenstates.

V. Role of the Systematic Structure in the Green's Function Matrix

Mathematical modeling is often most useful when it can identify features that allow for reductions in the complexity of the model without compromising its validity. In the case of the model problem from linear kinetics investigated in the previous section, the redundancy of the information content of the whole Green's function matrix is demonstrated by the reduction of the 144 (12×12) matrix elements to the 6 in Fig. 14 (and Table II). In quantum dynamics, similar considerations have lead to the formulation of scaling relations.¹⁷ The dramatic redundancy of Green's function matrix elements witnessed for the linear kinetics case, suggests something similar for the nonlinear kinetics as well, and is easily addressed by examining the gross structure of the whole Green's function matrix.

An examination of these simplifying features is particularly important for nonlinear kinetics since the "lumping" of complex models to obtain reduced pictures containing fewer parameters and variables is an important quest in the modeling of real engineering level kinetics problems. A knowledge of dynamic couplings between the various dependent variables can be a useful guide in this area and may help quantify the lumping of complex models which remains very much an art. Strong coupling between a set of dependent variables would imply that their response to any variation will be analogous and can be mimicked by retaining a single representative variable (or perhaps a special superposition of the dependent variables) from this set. In the previous sections, we examined the structure of the individual elements of the Green's function matrices from different problems to elucidate their role in the characterization of dynamic couplings between the dependent variables. In this section, we

examine systematic structure of the whole Green's function matrix and as a special example we again use the oxidation of carbon monoxide.

The Green's function coefficients of the pure temporal and the steady state reaction-convection-diffusion problems, obtained by solving Eqs. (III.6) and (III.7), respectively, for the reaction model described in Table I, form an 11×11 matrix (excluding the temperature). Some examples of individual Green's function surfaces from these problems have been offered previously and we have noted the evident similarities between the surfaces corresponding to different elements of the Green's function matrix. Specifically, Fig. 16 shows the surfaces for $\delta H_2O_2(t)/\delta J_O(t')$ and $\delta OH(t)/\delta J_H(t')$ for the temporal problem of Section III. We note that the two surfaces are nearly identical, although the magnitudes of their responses are different. This feature permeates the whole matrix of Green's function surfaces as evidenced in Fig. 17. In this figure, the elements represented by the same symbol have similarly behaved response surfaces and those with the same, but shaded, symbols are of opposite sign. An element without a symbol represents a response surface which could not be closely matched with the surface of any other element.

It is apparent from the systematic structure of this matrix that it can be conveniently partitioned between the major species and the intermediate species. This partitioning produces four non-square submatrices, each with their own characteristics. The elements of the intermediate species - intermediate species submatrix (lower right hand block) have similarly behaved response surfaces with the natural exception of the diagonal elements (i.e., the diagonal and off diagonal elements start out with distinctly different initial conditions). In contrast, the elements of the intermediate species-major species submatrix (lower left

hand block) are observed to have similarly behaved response surfaces for a given major species column. The element of the major species-intermediate species submatrix (upper right hand block) are observed to have only some elements with similarly behaved surfaces. Furthermore for this block, the similarities between the response surfaces occur along rows corresponding to major species. Finally, the elements of the major species - major species submatrix (upper left hand block) are observed to have the least similarity (blank spaces) among themselves. The present partitioning implies that all intermediate species respond in the same fashion and on the same time scale to variations in the flux of any intermediate species. Moreover, the major species respond in the same way to variations in any of the intermediate species but respond differently and in their own unique way to perturbations in other major species.

The maximum magnitudes of the elements in the lower two block matrices are $\sim 10^4$ larger than those in the upper block matrices. However, if the coefficients are logarithmically normalized (i.e., multiplied by O_j/O_1), this significant difference is practically eliminated since the major species concentrations are generally larger than the intermediate species concentrations by $\sim 10^3$. Also, the Green's function matrix elements involving molecular hydrogen react under some circumstances as if molecular hydrogen were a major species (e.g., when its concentration is perturbed) and at other times as an intermediate species.

The Green's function matrix for the steady premixed flame $\text{CO}/\text{H}_2/\text{O}_2$ oxidation problem, is pictorially shown in Fig. 18. Once again, similarities of the kind discussed for the pure temporal problem can be observed between the various elements of the system. It is interesting to compare the structure of this matrix with the matrix obtained from the

temporal problem. While obvious similarities exist among the intermediates, just as in the pure temporal case, some distinct differences such as the separation of behavior of the heavier intermediates HO_2 and H_2O_2 from the lighter intermediates O , H , OH and HCO , due to diffusion effects, is readily apparent.

Pronounced similarities amongst elements of the Green's function matrix of the kind found above have been observed in other problems as well⁸ and have prompted the search for unifying relations to be discussed below.¹⁸ The reason for such similarities is best explored in the context of the steady laminar flame problem. In this case, the similarity of various response surfaces to each other and to the temperature response surfaces is associated with the dominant role of the temperature in combustion problems, as suggested by its more extreme nonlinear role in comparison to the chemical species. The presence of a dominant variable, i.e., temperature, leads to scaling and self similarity relations between dependent variables and the topic has been treated in detail elsewhere.¹⁸ Recent work has also shown that the presence of significant diffusion can enhance the presence of scaling and self similarity.¹⁹ These relations can explain the similar details of Green's function surfaces in the examples cited above. Though in the case of pure temporal isothermal kinetics, no single dominant variable is easily identified, an extension of the same analysis can be brought to bear on the problem.¹⁸ A brief synopsis of scaling and self similarity results is given below.

The scaling and self similarity relations ensue from the simple ansatz of the presence of a single dominant variable (to be denoted by O_1). As a result of this assumption, we may separate Eq. (II.1) into two parts

$$L_1(O, \underline{\alpha}) = 0 \quad (V.1)$$

$$L_i(O, \underline{\alpha}) = 0 \quad ; \quad i = 2, 3, \dots, N \quad (V.2)$$

The dominant role of O_1 is manifest in the conjecture

$$O_n(x, \underline{\alpha}) \in F_n[O_1(x, \underline{\alpha})] \quad (V.3)$$

that the x and $\underline{\alpha}$ dependence of $O_n(x, \underline{\alpha})$ essentially arises as a function F_n of the dependence occurring in the dominant controlling dependent variable $O_1(x, \underline{\alpha})$.

Functional differentiation of (V.3), with respect to $J_n(x')$ leads to

$$\left(\frac{\delta O_n(x)}{\delta J_n'(x')} \right) = \left(\frac{\partial F_n}{\partial O_1} \right) \left(\frac{\partial O_1}{\delta J_n'(x')} \right) \quad (V.4)$$

and similarly, differentiating Eq. (V.3) with respect to x results in

$$\left(\frac{\partial O_n}{\partial x} \right) = \left(\frac{\partial F_n}{\partial O_1} \right) \left(\frac{\partial O_1}{\partial x} \right) \quad (V.5)$$

Eliminating the derivative $\partial F_n / \partial O_1$ from Eqs. (V.4) and (V.5) we obtain the scaling relation

$$\left(\frac{\delta O_n(x)}{\delta J_n'(x')} \right) = \left(\frac{\delta O_1(x)}{\delta J_n'(x')} \right) \left(\frac{\partial O_n}{\partial x} \right) \left(\frac{\partial O_1}{\partial x} \right)^{-1} \quad (V.6)$$

This equation implies that all the elements of the Green's function matrix may be deduced from the first column of that matrix in conjunction with a knowledge of the coordinate gradients of the corresponding dependent variables. The scaling implied by Eq. (V.6) corresponds to a reduction of the $N \times N$ dimensional Green's function matrix down to knowledge of a single vector of dimension N . An immediate consequence of Eq. (V.6) is the relations

$$\frac{(\delta O_n(x)/\delta J_k(x'))}{(\delta O_n(x)/\delta J_{k'}(x'))} \approx \frac{(\delta O_1(x)/\delta J_k(x'))}{(\delta O_1(x)/\delta J_{k'}(x'))} \quad (V.7a)$$

$$\frac{(\delta O_n(x)/\delta J_k(x'))}{(\delta O_{n'}(x)/\delta J_k(x'))} \approx \frac{(\delta O_n(x)/\delta x)}{(\delta O_{n'}(x)/\delta x)} \quad (V.7b)$$

The simplification implied by these results is quite dramatic and their validity is easily tested. For example, a simple consequence is that the Green's function elements of the n^{th} dependent variable will change sign as a function of x whenever an extrema $(\delta O_n/\delta x) = 0$ exists (Fig. 19a has this behavior upon examination of $\partial H(x)/\partial x$ (not shown here)). In addition, manipulation of Eq. (V.7b) will show that it has the same structure as Eq. (V.6), except now the dominant role is replaced by $O_{n'}$ as an arbitrary member of the strongly coupled set of dependent variables. These results suggest that the choice of O_1 is dominant in Eq. (V.3), may be relaxed to any member of the strongly coupled set of dependent variables. This statement is also supported by numerical evidence validating Eq. (V.7).^{18,19}

In steady state flame problems, the temperature is a monotonically increasing function of x , with positive slope, and the monotonically decreasing reactant concentration will have a negative slope. As a consequence, with the identification of temperature as the dominant variable, we can understand that the Green's function surfaces for the reactants are basically the negative of the corresponding Green's functions for the temperature as seen from comparing Figs. 19b and 19c. Since, due to the conservation of mass, a decrease in CO would always lead to an increase in CO₂ concentration (the HCO concentration is inconsequential), it also explains why the columns corresponding to CO and CO₂ are the

obverse of each other (see Fig. 18). The structure of the Green's function surfaces for intermediates may be similarly understood. The intermediate concentration, which is initially zero at the inlet, rises to a maximum in the flame zone and then decreases. As a consequence, the intermediate Green's function matrices should look similar to the temperature Green's function but change sign upon passage through the flame as exemplified by Fig. 19a. The similarities between the surfaces $\delta CO(x)/\delta J_H(x')$ and $\delta CO(x)/\delta J_{OH}(x')$ (Fig. 3) is easily explained by Eq. (V.7a) and the preceding discussion regarding the response functions for the intermediates.

While the role of the scaling relation Eq. (V.6) as an organizing principle is made plain by the examples cited above, the use of Eq. (V.6) in conjunction with Eq. (III.7) leads to further simplifications. The substitution of Eq. (V.6) into (III.7) ultimately leads to the following result^{18,19}

$$\frac{\delta O_1(x)}{\delta J_{n'}(x')} \approx \lambda(x) \begin{cases} f_{n'}^+(x') & x > x' \\ f_{n'}^-(x') & x < x' \end{cases} \quad (V.8)$$

where $\lambda(x)$ and $f_{n'}^\pm$ are system characteristic functions. The validity of this construct is borne out by Fig. 4, where the requisite discontinuity $f_{n'}^-(x') \neq f_{n'}^+(x')$ at $x=x'$ and the factorization of the Green's function according to Eq. (V.8) is apparent. This feature persists in other surfaces as well with different levels of smoothness in the jump across $x=x'$.

Finally, substitution of Eq. (V.8) into Eq. (V.6) leads to the self similarity relation

$$\frac{\delta O_n(x)}{\delta J_{n'}(x')} \approx \Lambda_n(x) \begin{cases} f_{n'}^+(x') & x > x' \\ f_{n'}^-(x') & x < x' \end{cases} \quad (V.9)$$

$$\Lambda_n(x) = \lambda(x) \left(\frac{\partial O_n}{\partial x} \right) \left(\frac{\partial O_1}{\partial x} \right)^{-1} \quad (V.10)$$

We note that the Eqs. (V.6) and (V.10) imply that the N^2 Green's function surfaces are completely characterized by the N dimensional vector of surfaces $[\delta O_1(x)/\delta J_{n'}(x')]$, which themselves are a simple product of functions indicated in Eq. (V.8). The physical significance of Eq. (V.9) is evident when we recall that these response functions are measurable in the laboratory. In particular, taking $n=1$, the function $\lambda(x)$ may be determined by disturbing the flux of any dependent variable and the functions $f_{n'}^+(x')$ could be determined by disturbing the flux of each of the dependent variables in turn. The scaling and self similarity relations, therefore, offer insight into the structure of response surfaces and make plain the nature and extent of dynamic couplings. The general conditions for the validity of the scaling and self similarity relations still needs to be firmly established. However, computational evidence suggests that relations become increasingly valid in the presence of strong dynamical coupling, regardless of whether its origin is through kinetics, diffusion or thermal effects.

VI. Concluding Remarks

We have attempted to illustrate the role of Green's functions in physically characterizing the dynamic couplings in diverse chemical/physical phenomena. The nonlinearity in chemical kinetics and the

interference structure in quantum dynamics lead to effects that transcend the apparent network structure between the dependent variables of the underlying models. The elements of the Green's function matrix help elicit the nature and extent of dynamic couplings between the dependent variables of a model system. While the coupling or its absence between any two dependent variables is revealed by the corresponding element of the Green's function matrix, the possible role of other dependent variables in mediating this coupling may only be ascertained by freezing appropriate variables selectively, and using the concept of reduced Green's functions. In this paper, we have illustrated the interpretive utility of Green's functions by offering examples of their use in pure temporal kinetics, reacting-diffusing-convecting steady state kinetics, and from some model problems in quantum dynamics.

The possibility of reducing the complexity of any mathematical model can depend on the ability to identify a set of dependent variables with similar response to various system parameters. Such an identification may make possible the use of a reduced number or a single representative member from this set for effective modeling of the system behavior. A global characterization of the Green's function matrix, exemplified by a block structure or reduced rank, suggests such a possibility. At least in the case of some kinetics problems, the presence of scaling and self similarity relations directly implies a reduced rank for the Green's function matrix. The ease with which Green's functions yield insights into dynamic system couplings in diverse chemical/physical systems augurs well for their wider application in the future.

ACKNOWLEDGMENTS The authors acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

1. R. P. Feynman and A. R. Hibbs, Quantum Mechanics and Path Integrals, (McGraw Hill:New York, 1965).
2. J. Linderberg and Y. Ohrn, Propagators in Quantum Chemistry, (Academic Press:New York, 1973).
3. H. Rabitz, M. Kramer and D. Dacol, Ann. Rev. Phys. Chem. **34**, 419 (1983); H. Rabitz, Computers and Chemistry **5**, 167 (1981).
4. M. Demiralp and H. Rabitz, J. Chem. Phys. **74**, 3362 (1981).
5. J. T. Hwang and H. Rabitz, J. Chem. Phys. **70**, 4609 (1979).
6. R. Hedges and H. Rabitz, J. Chem. Phys. **82**, 3674 (1985).
7. I. Gumoski, in Sensitivity Methods in Control Theory, (L. Radnovic, ed., Pergamon Press:Oxford, 1966).
8. Y. Reuven, M. D. Smooke and H. Rabitz, J. Comp. Phys. **64**, 27 (1986).
9. M. D. Smooke, J. Comp. Phys. **48**, 72 (1982).
10. R. Yetter, F.L. Dryer and H. Rabitz, Fall Western States Section Meeting, The Combustion Institute, WSS paper #84-86, 1984. An updated mechanism is available from R. Yetter, F. L. Dryer and H. Rabitz, submitted to Combust. Sci. Tech, 1989.
11. A. C. Hindmarsh, in ACM Signum Newsletter **15**(4), (1980).
12. M. A. Kramer, J. M. Calo, H. Rabitz, and R. J. Kee, Sandia Technical Report 82-9231, Sandia National Laboratory, Livermore, CA, 1982.
13. R. Yetter, F. L. Dryer and H. Rabitz, Combust. Flame **59**, 107 (1985).
14. M. Mishra, R. Yetter, Y. Reuven, H. Rabitz, and M. D. Smooke, *"Sensitivity Analysis of Steady State Premixed Laminar Flames Using Newton's Method: Application to the CO+H₂+O₂ System"*, to be published.
15. A. Messiah, Quantum Mechanics, volume II, (North Holland:Amsterdam, 1976, p. 722).

16. K. S. J. Nordholm and S. A. Rice, J. Chem. Phys. 61, 203 (1974).
17. A. E. Depristo and H. Rabitz, J. Chem. Phys. 68, 1981 (1978).
18. H. Rabitz and M. D. Smooke, J. Phys. Chem. 92, 1110 (1988).
19. S. Vajda, R. Yetter and H. Rabitz, Combustion and Flame, in press.

Figure Captions

1. Response surface for the Green's function elements (a) $\delta\text{CO}(t)/\delta J_H(t')$ and (b) $\delta\text{CO}(t)/\delta J_{OH}(t')$ for the pure temporal system of Table I. The peak response of Fig. 1a is approximately twice that of Fig. 1b.
2. Cuts (at $t'=0$) of the reduced Green's functions $\delta\text{CO}(t)/\delta J_x(t')|_{\delta OH=0}$ ($x=H, O$, etc.) for the pure temporal problem where the OH species have been frozen. These responses are dramatically smaller (by 4-5 orders of magnitude) than those with unconstrained OH thus revealing the critical role of OH in the oxidation of CO.
3. The response surface for a premixed steady $\text{CO}/\text{H}_2/\text{O}_2$ laminar flame. a) $\delta\text{CO}(x)/\delta J_H(x')$ and b) $\delta\text{CO}(x)/\delta J_{OH}(x')$. The maximum of the response to the perturbation in the flux of H atoms is about 20 times more than that due to the perturbation in the OH flux. This is easily understood due to the greater mobility of the lighter species H.
4. The response surface for the reduced Green's function $\delta\text{CO}(x)/\delta J_H(x')|_{\delta OH=0}$ of the laminar flame problem. Not only is the magnitude of the response reduced by a factor of twenty in comparison with that in Fig. 3, the freezing of OH response also leads to a reversal in the role of H atoms. A small added flux of H atoms is seen now to inhibit the consumption of CO.
5. Model Hamiltonian for two coupled oscillators with two and eight eigenstates, respectively. The nondiagonal elements determine the initial kinematic coupling structure between the eigenstates of the unperturbed Hamiltonian. The heavy bold lines through the matrix

separate the two blocks of eight states involving, respectively, first and second eigenstates of the second oscillator. Shading scale to the right of the figure corresponds to the numerical magnitude of the matrix elements in the Hamiltonian.

6. The long term mean square average of the Green's function corresponding to the Hamiltonian in Figure 5. The shading scale to the right of the figure corresponds to the magnitude of the matrix elements. The simple nearest neighbor coupling structure of the Hamiltonian leads to a long time behavior where almost all the states are strongly coupled to each other.
7. A variation on the system represented in Fig. 5. Here, both oscillators have four states, and the groupings are denoted by the bold lines. The shading scale of Fig. 5 applies here.
8. The long time average of the Green's function for the Hamiltonian in Fig. 7. The shading scale of Fig. 6 applies here. The sharply banded structure shows that the energy distribution of dynamic coupling is neither limited to the originally coupled states alone nor is it entirely random.
9. Matrix representing a pentadiagonal Hamiltonian. The shading scale of Fig. 5 applies here. The differences in the magnitude of the nondiagonal elements mimic a varied kinematic coupling structure.
10. The long time mean square average of the Green's function for the Hamiltonian in Fig. 9. The shading scale of Fig. 6 applies here. The marked difference between the structure of the Hamiltonian and the long time coupling between the states underscores the dynamical content of the Green's functions.

11. The long time average of the reduced Green's function for the Hamiltonian in Fig. 9, where the 7th state has been eliminated. The shading scale of Fig. 6 applies here. The block diagonal nature of the reduced Green's function matrix reveals the critical role of state 7 as a gateway for coupling between states 1-6 and 8-16.
12. The long time average of the reduced Green's function for the Hamiltonian in Fig. 9 where the 9th eigenstate has been eliminated. The shading scale of Fig. 6 applies here. The elimination of this state increases the transition probability between states 11 and 12, revealing its role as a bottleneck for dynamic coupling between these two states. Aside from this change, the state 9 has a much less critical role than that of state 7 which is apparent from a comparison of Figs. 10, 11 and 13.
13. The matrix which doubles as both $\underline{\underline{H}}$ and $\underline{\underline{M}}$. The conservation of matter in kinetics necessitates that the diagonal element in any column equal the negative of the sum of the remaining positive off diagonal elements of that column. The real symmetric nature of the matrix permits its use to represent a Hamiltonian operator $\underline{\underline{H}}$ as well.
14. Green's function matrix elements from the linear kinetics problem with $\underline{\underline{M}}$ represented in Fig. 13. The linearity of the system leads to an entirely predictable behavior with the magnitude of response being determined by the closeness of coupling $|i-j|$. The curves a-,f correspond to particular elements shown in Table II.
15. The long time mean square average quantum mechanics Green's function for the Hamiltonian in Fig. 13. The shading scale of Fig. 6 applies here. Unlike linear kinetics with the same matrix $\underline{\underline{M}} = \underline{\underline{H}}$ in Fig. 14,

the interference structure in quantum mechanics leads to nearly uniform coupling of all states.

16. Comparison of the Green's function coefficient surfaces for the elements, (a) $\delta H_2O_2(t)/\delta J_O(t')$ and (b) $\delta OH(t)/\delta J_H(t')$ from the temporal, isothermal wet oxidation of CO. The maximum and minimum values for $\delta H_2O_2(t)/\delta J_O(t')$ are 798 and -81, respectively. The corresponding values for $\delta OH(t)/\delta J_H(t')$ are 133 and -17.
17. Schematic diagram of the wet CO oxidation Green's functions $\delta \sim_i(t)/\delta J_j(t')$ from the temporal problem. Elements of similar shape have similarly behaved Green's function surfaces as a function of t and t' . Those with the same shape, but shaded in, are of opposite sign. Blank spaces indicate a response surface which could not be closely matched with another.
18. Schematic diagram of the steady CO/H₂/O₂ premixed flame Green's functions, $\delta \sim_i(x)/\delta J_j(x')$. The conventions of Fig. 17 apply.
19. The response surfaces from the premixed CO/H₂/O₂ laminar flame problem: (a) $\delta H(x)/\delta J_O(x')$, (b) $\delta CO(x)/\delta J_O(x')$ and (c) $\delta T(x)/\delta J_O(x')$. These figures with their similar structures illustrate the scaling and self similarity relations in Section V.

Table I. CO/H₂/O₂ Kinetic Mechanism

No.	Reaction	A ¹	n	E	I ²	UF ³
1,2 ³	HCO + H = CO + H ₂	2.00(14) ⁴	0.0	0.0	f	2
3,4	HCO + OH = CO + H ₂ O	1.00(14)	0.0	0.0	f	3.5
5,6	O + HCO = CO + OH	3.02(13)	0.0	0.0	f	2
7,8	HCO + O ₂ = CO + HO ₂	3.01(12)	0.0	0.0	f	1.5
9,10	CO + HO ₂ = CO ₂ + OH	1.50(14)	0.0	2.36(4)	f	2
11,12	CO + OH = H + CO ₂	4.46(6)	1.5	-7.40(2)	f	1.5
13,14	CO ₂ + O = CO + O ₂	2.53(12)	0.0	4.77(4)	b	3
15,16	H + O ₂ = O + OH	3.73(17)	-1.0	1.75(4)	f	2
17,18	H ₂ + O = H + OH	1.80(10)	1.0	8.90(3)	f	2
19,20	O + H ₂ O = OH + OH	4.58(9)	1.3	1.71(4)	f	2.5
21,22	H + H ₂ O = OH + H ₂	1.08(9)	1.3	3.65(3)	b	2
23,24	H ₂ O ₂ + OH = H ₂ O + HO ₂	7.00(12)	0.0	1.43(3)	f	2
25,26	HO ₂ + O = O ₂ + OH	1.81(13)	0.0	-3.97(2)	f	2
27,28	H + HO ₂ = OH + OH	1.69(14)	0.0	8.74(2)	f	1.5
29,30	H + HO ₂ = H ₂ + O ₂	6.63(13)	0.0	2.13(3)	f	2
31,32	OH + HO ₂ = H ₂ O + O ₂	1.45(16)	-1.0	0.0	f	2.5
33,34	H ₂ O ₂ + O ₂ = HO ₂ + HO ₂	1.00(13)	0.0	1.00(3)	b	3
35,36	HO ₂ + H ₂ = H ₂ O ₂ + H	1.70(12)	0.0	3.75(3)	b	2
37,38	O ₂ + M = O + O + M	6.17(15)	-0.5	0.0	b	3
39,40	H ₂ + M = H + H + M	2.20(14)	0.0	9.60(4)	f	2
41,42	OH + M = O + H + M	1.00(16)	0.0	0.0	b	30
43,44	H ₂ O ₂ + M = OH + OH + M	1.20(17)	0.0	4.55(4)	f	2
45,46	H ₂ O + M = H + OH + M	2.20(16)	0.0	1.05(5)	f	2
47,48	HO ₂ + M = H + O ₂ + M	1.65(15)	0.0	-1.00(3)	b	3

49,50	$\text{CO}_2 + \text{M} = \text{CO} + \text{O} + \text{M}$	5.90(15)	0.0	4.10(3)	b	4
51,52	$\text{HCO} + \text{M} = \text{H} + \text{CO} + \text{M}$	6.90(14)	0.0	7.00(3)	b	1.5
53,54	$\text{H} + \text{H}_2\text{O}_2 = \text{H}_2\text{O} + \text{OH}$	1.00(13)	0.0	3.59(3)	f	3

$[\text{M}] = [\text{N}_2] + [\text{O}_2] + 16[\text{H}_2\text{O}] + 2.5[\text{H}_2] + 3.8[\text{CO}_2] + 1.9[\text{CO}] + [\text{HO}_2] + [\text{H}_2\text{O}_2] + [\text{H}] +$
 $[\text{O}] + [\text{OH}] + [\text{HCO}] + 0.87[\text{Ar}]$

¹ Units are cm-mole-sec-cal, $k = AT^n \exp(-E/RT)$

² I indicates direction of reaction for which rate constant data was used.

References for the rate data may be found in Refs. 10 and 13.

³ Number associated with forward rate constant, number associated with reverse rate constant.

⁴ Numbers in parentheses denote powers of ten.

Table II:

A tabulation of the Green's function matrix elements for the linear kinetics system described in Section IV and Fig. 14. Due to the symmetric nature of the matrix, only the lower triangle portion is displayed. Elements represented by the same letter have an identical response profile corresponding to the curves in Fig. 24.

	1	2	3	4	5	6	7	8	9	10	11	12
1	a											
2	b	a										
3	c	b	a									
4	d	c	b	a								
5	e	d	b	b	a							
6	f	e	d	c	b	a						
7	f	f	e	d	c	b	a					
8	f	f	f	e	d	c	b	a				
9	e	f	f	f	e	d	c	b	a			
10	d	e	f	f	f	e	d	c	b	a		
11	c	d	e	f	f	f	e	d	c	b	a	
12	b	c	d	e	f	f	f	e	d	c	b	a

Figure 1

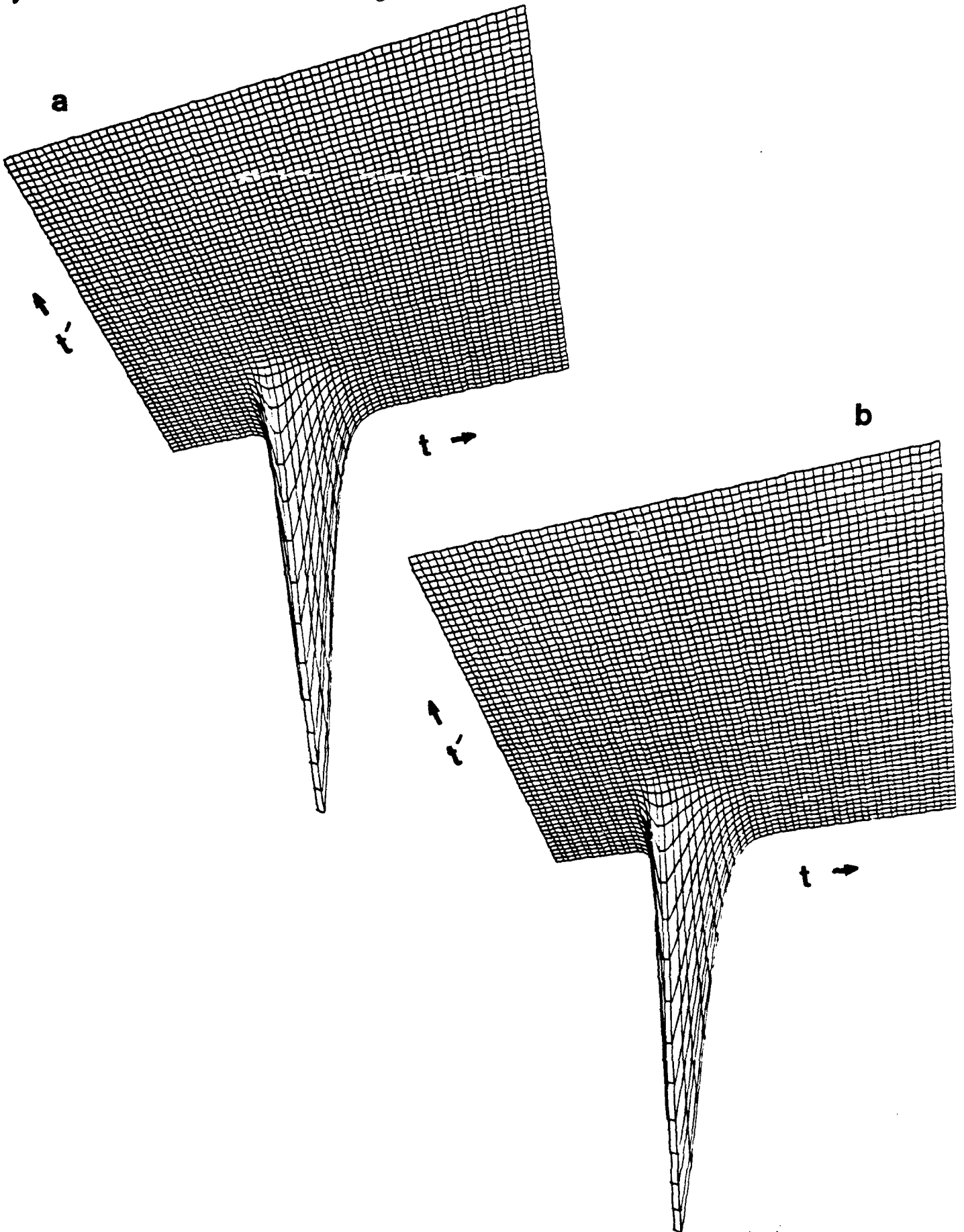


Figure 2

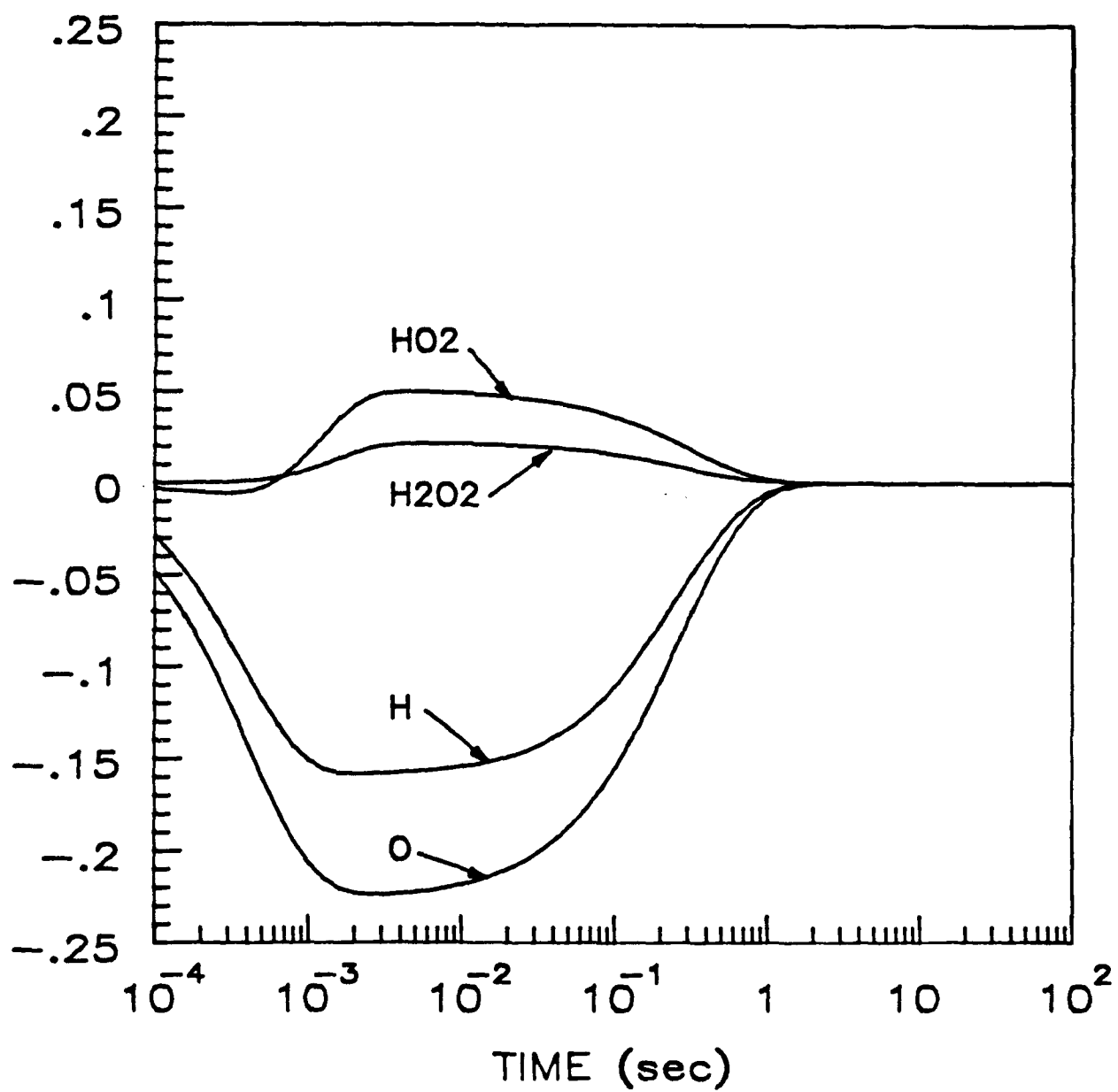


Figure 3

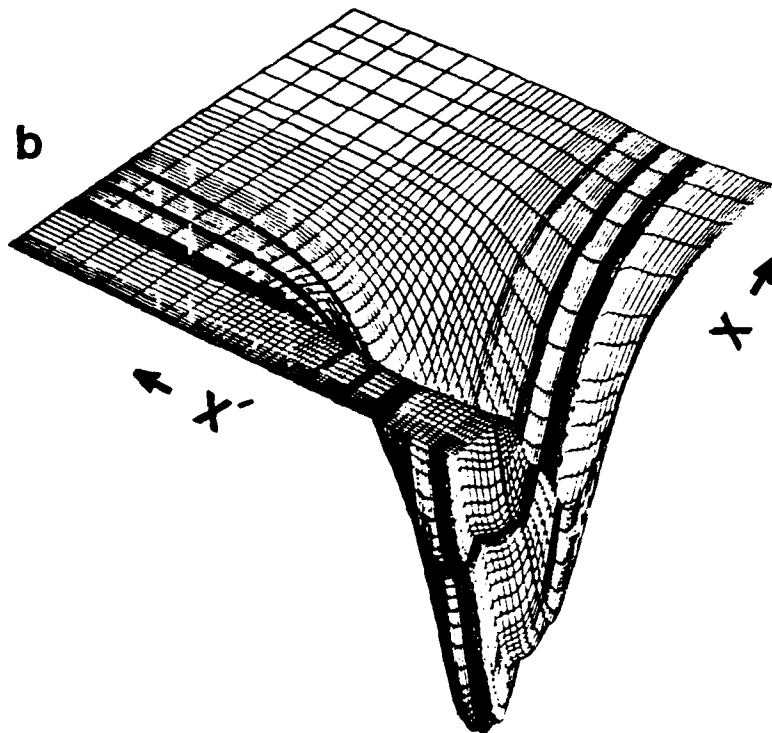
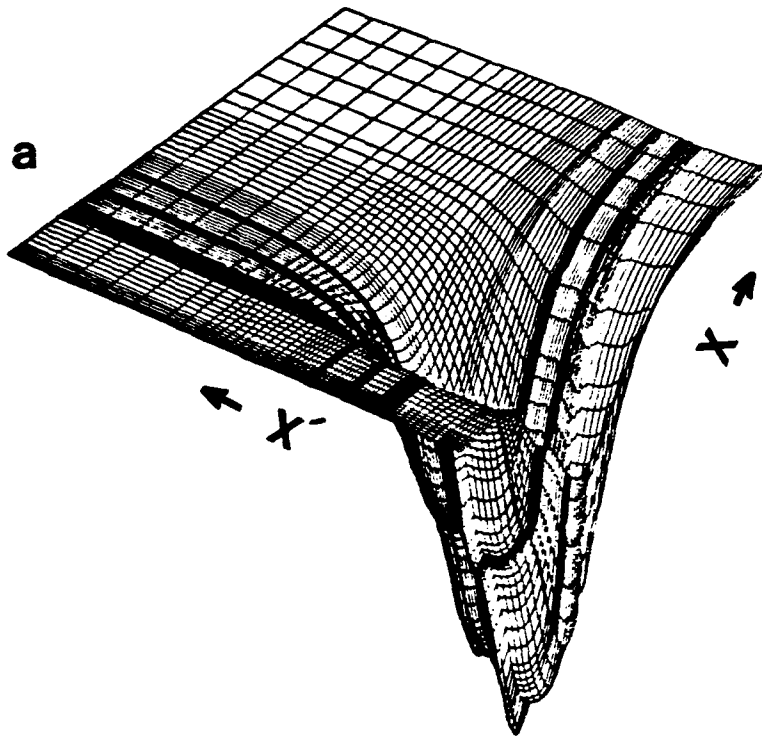
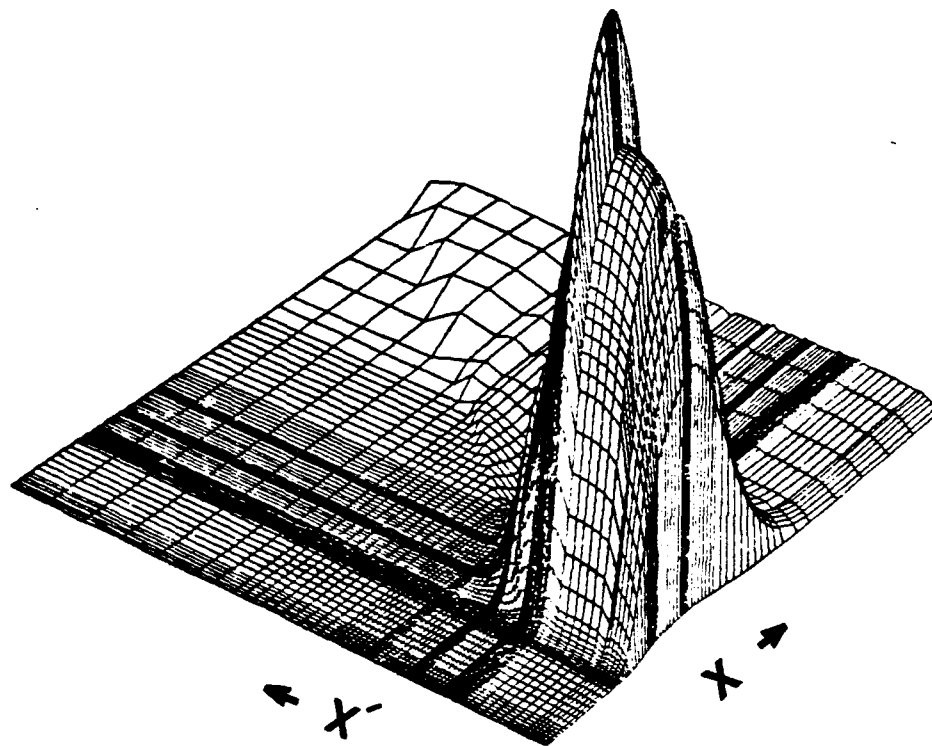


Figure 4



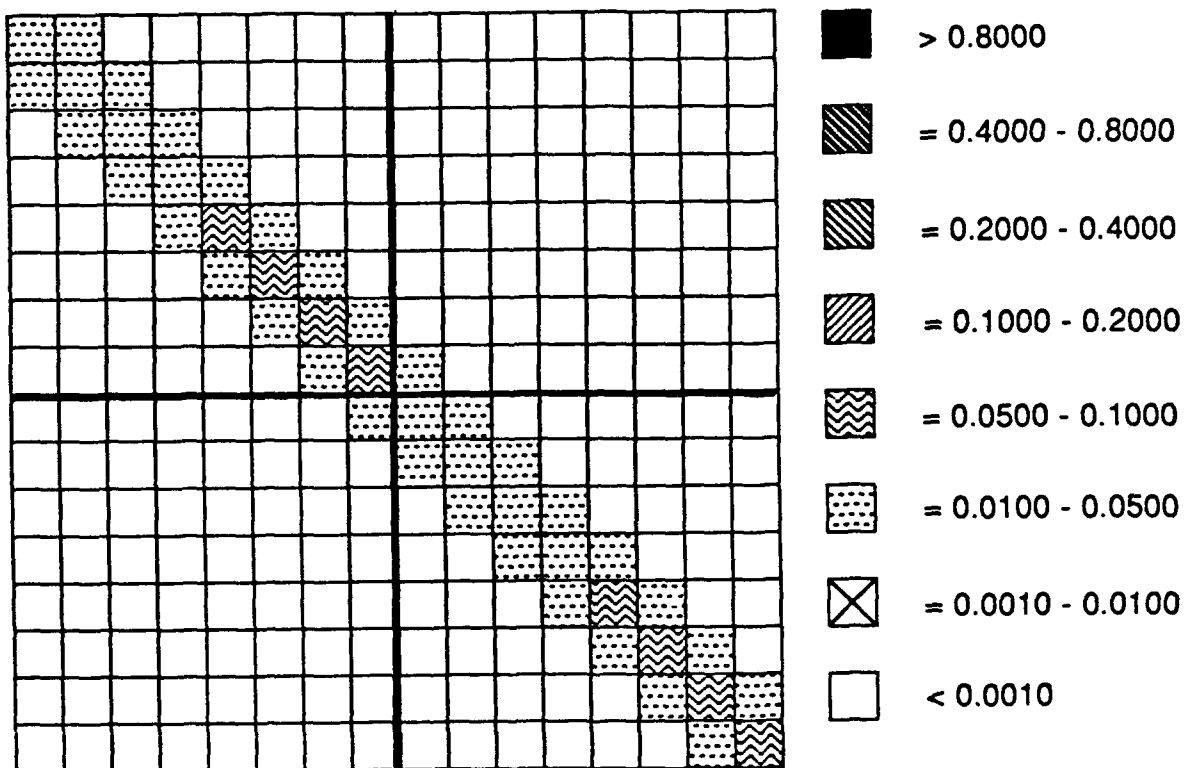


Figure 5

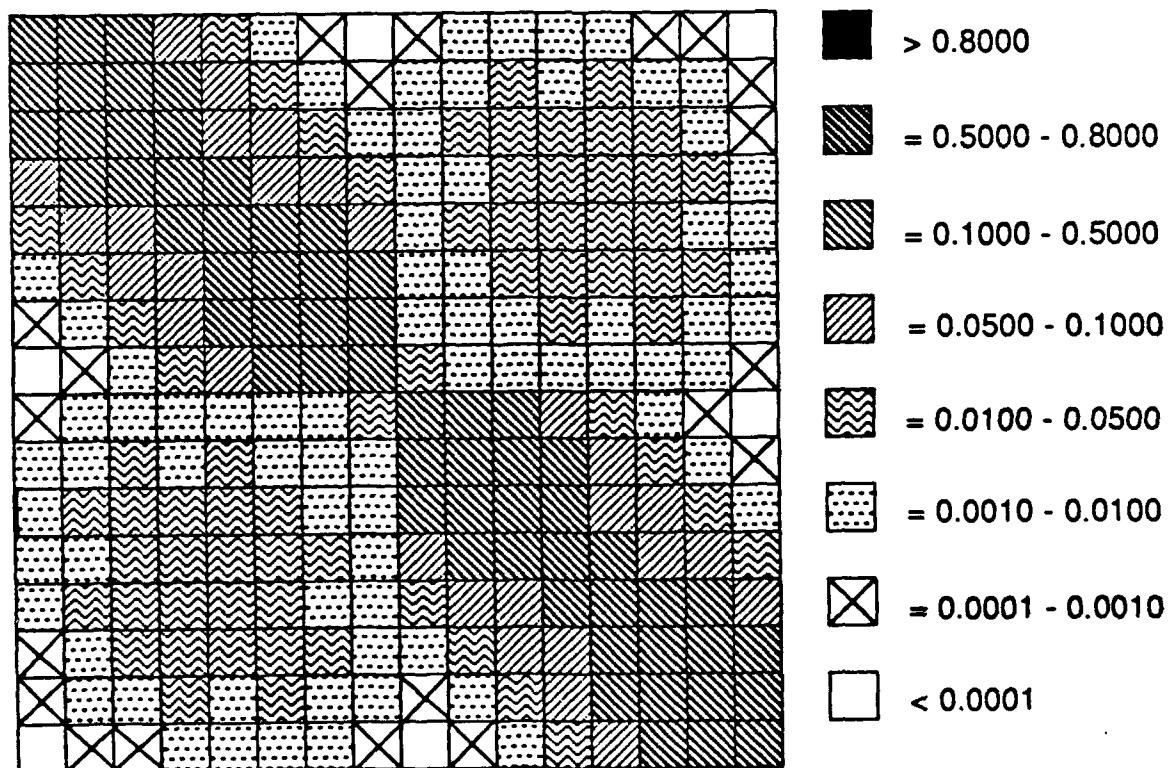


Figure 6

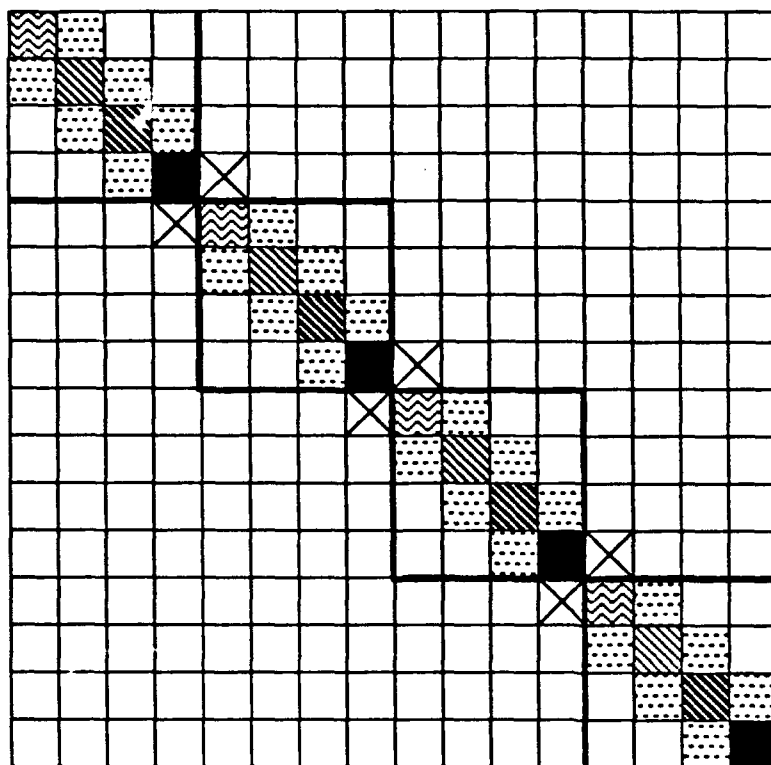


Figure 7

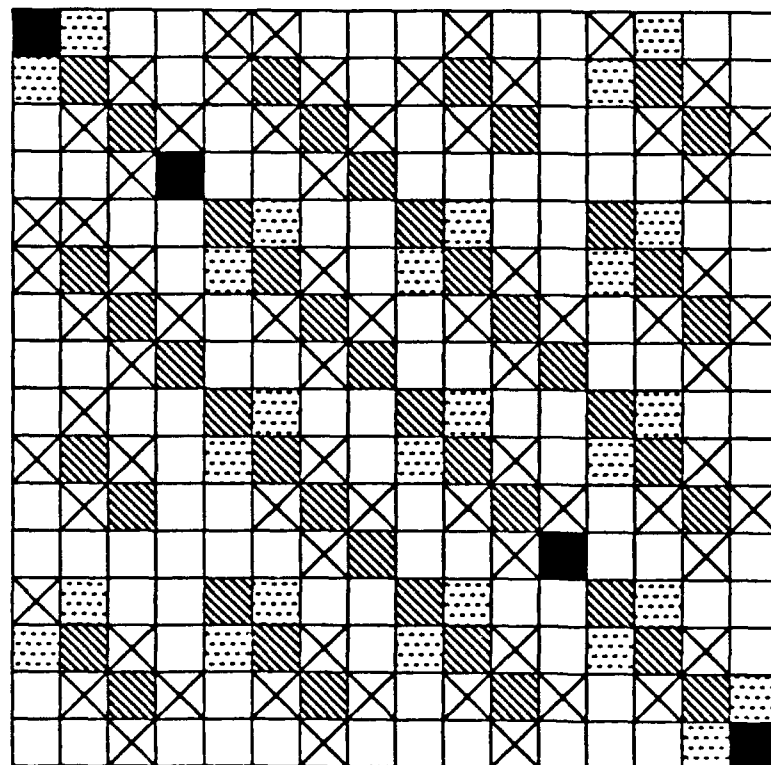


Figure 8

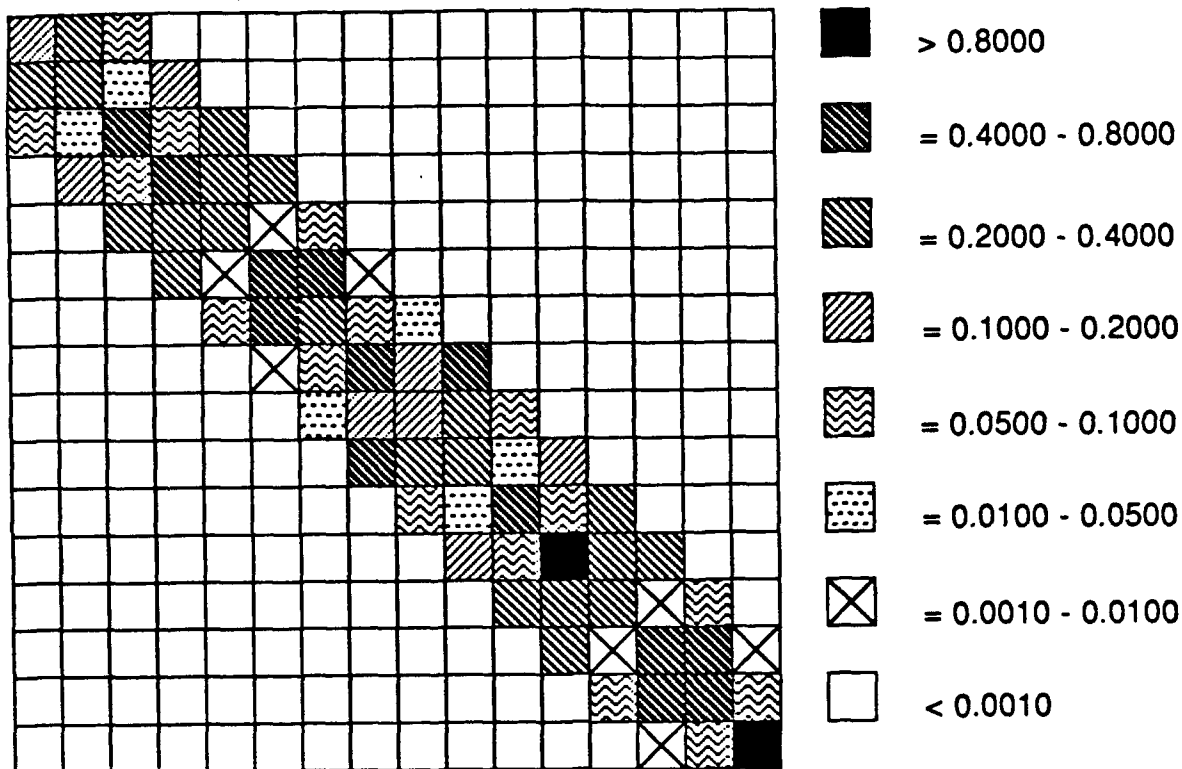


Figure 9

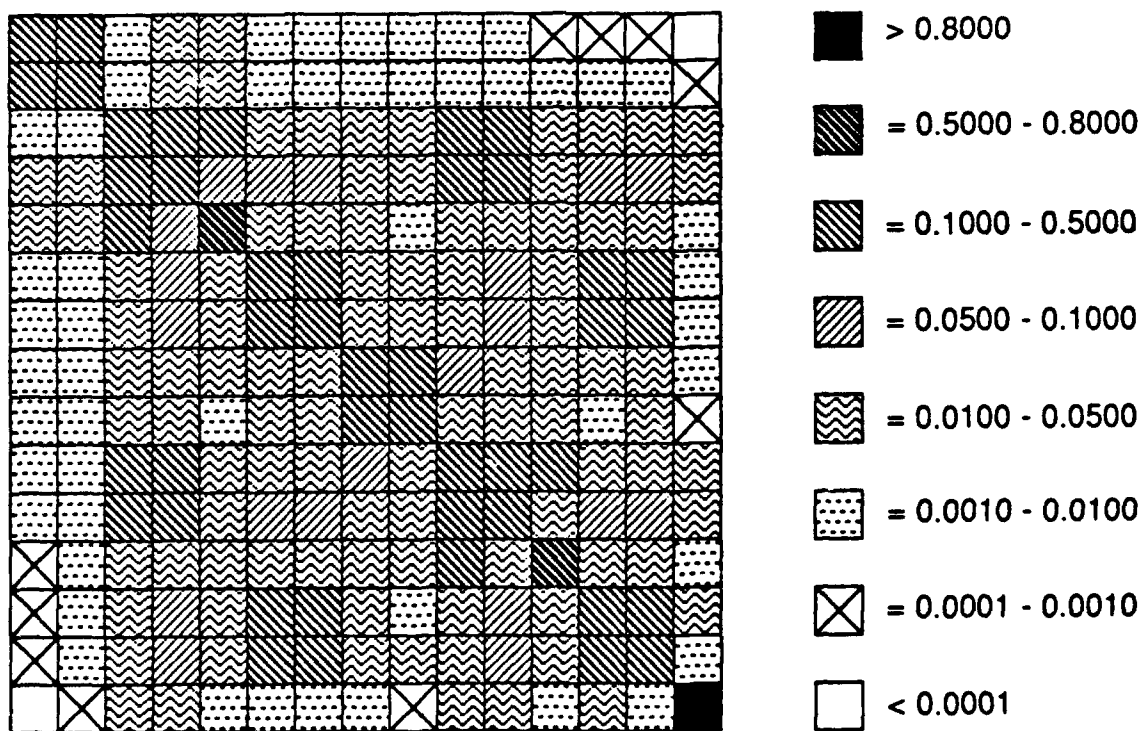


Figure 10

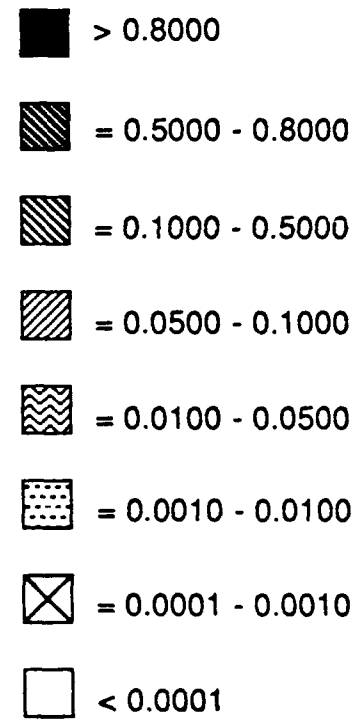
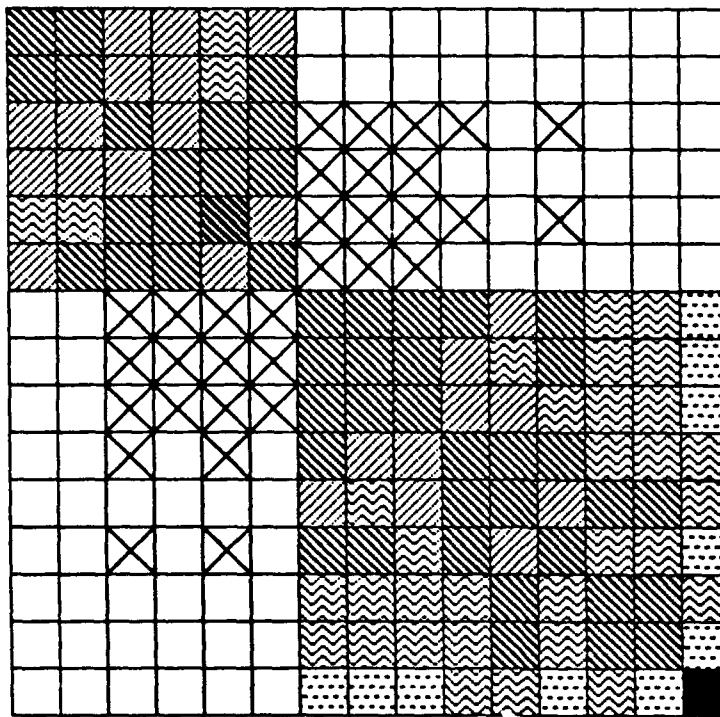


Figure 11

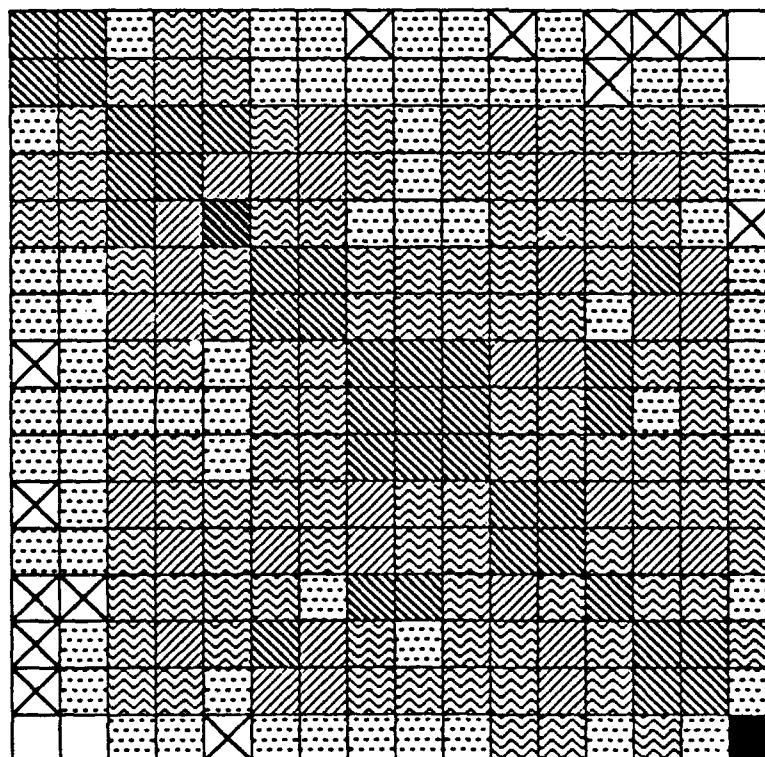


Figure 12

Figure 13

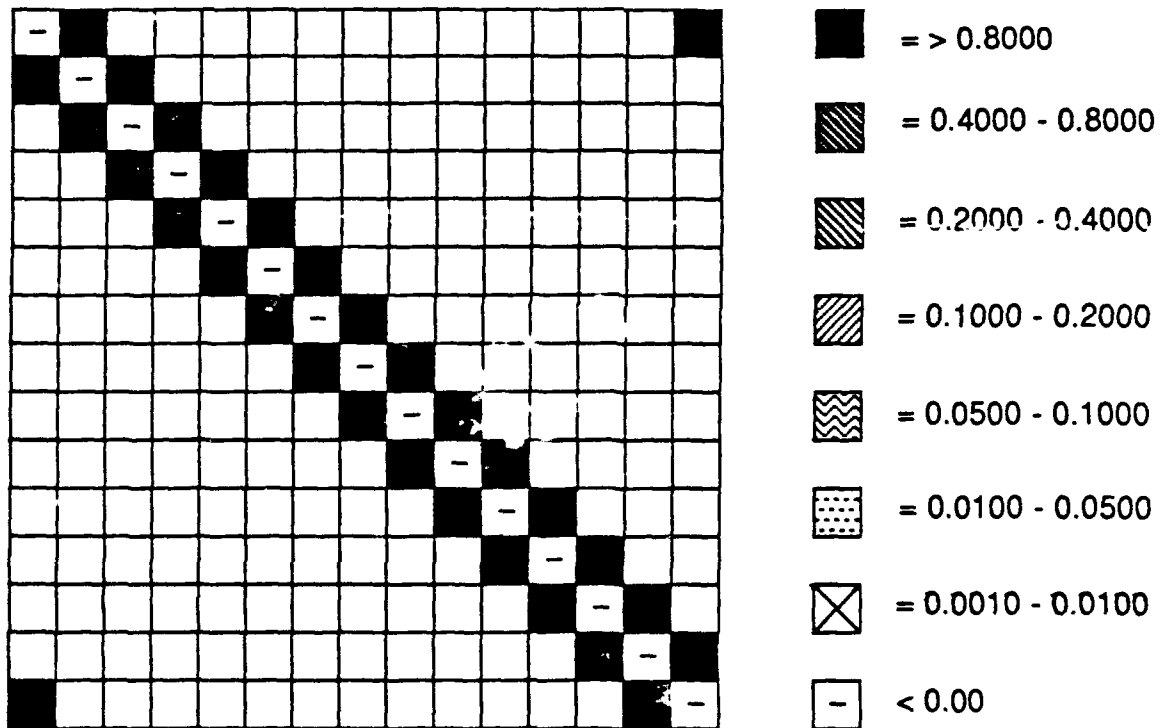


Figure 14

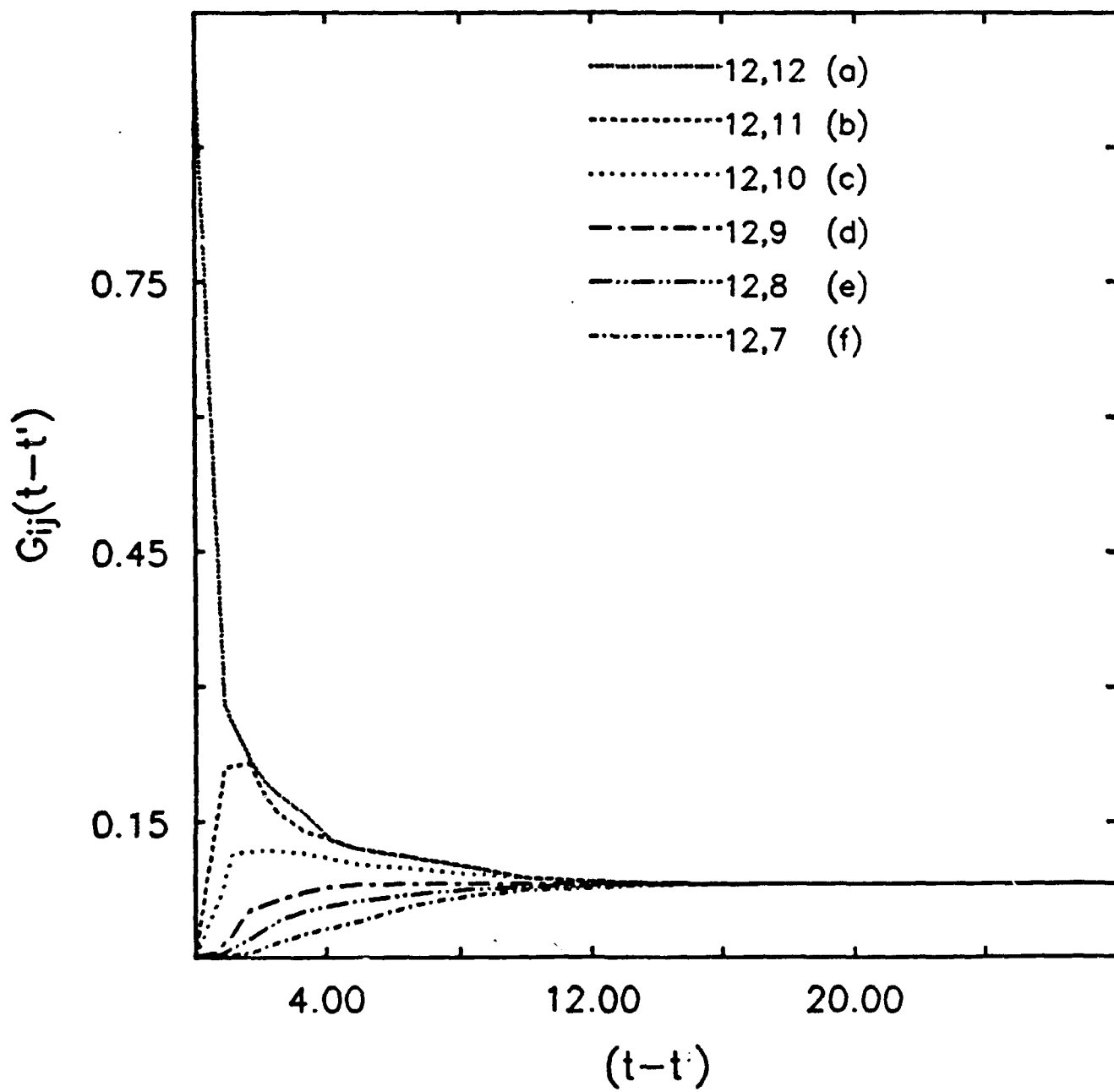


Figure 15

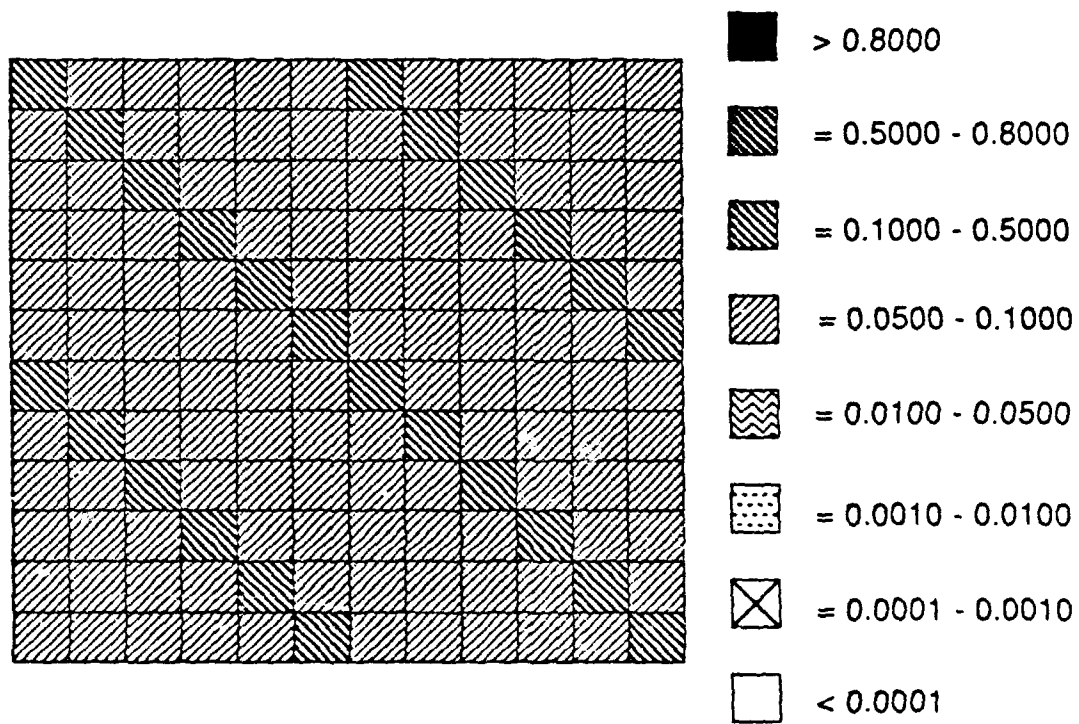


Figure 16

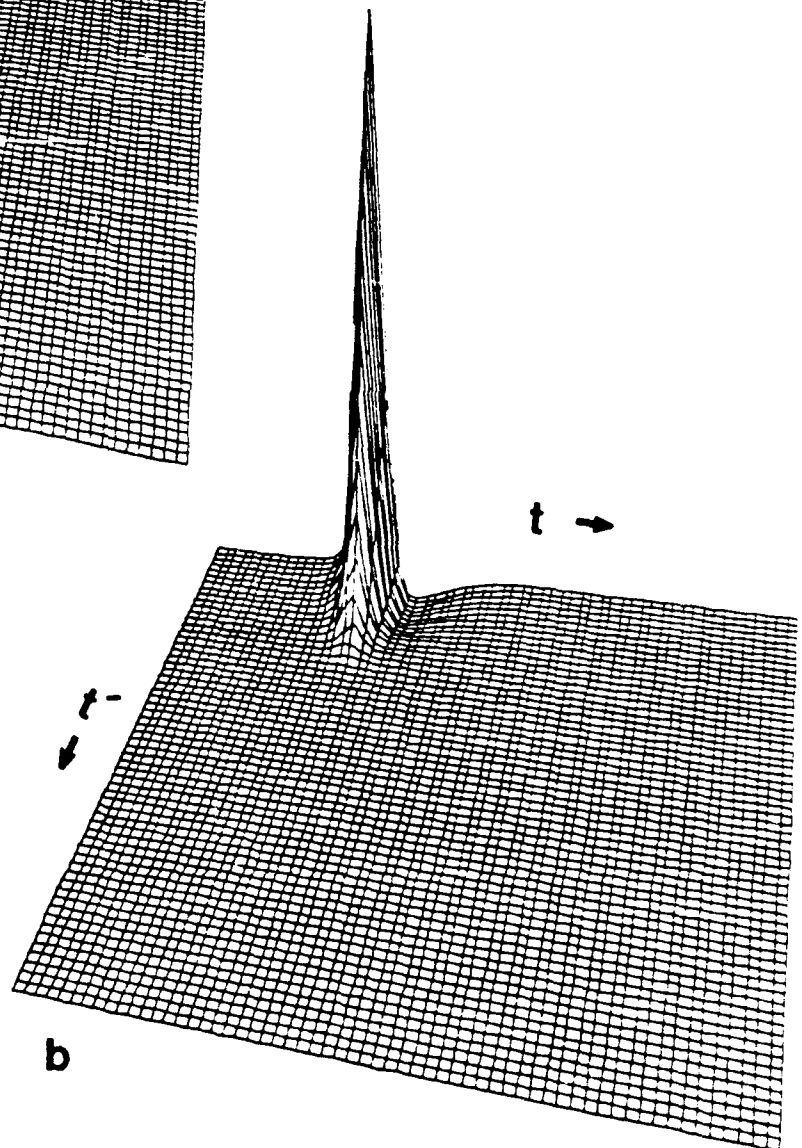
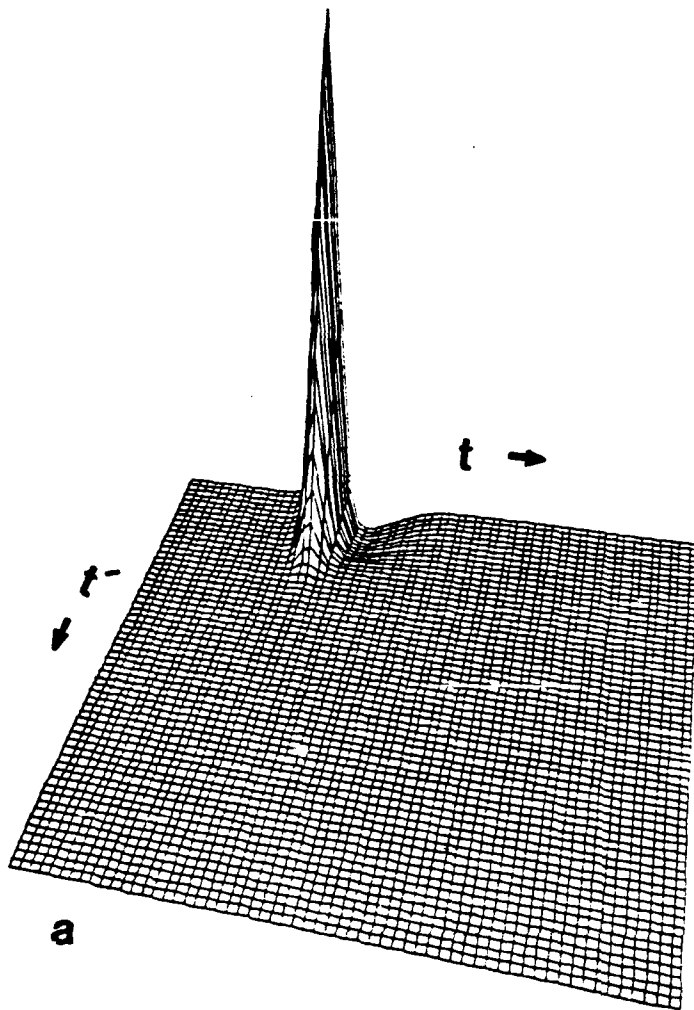


Figure 17

































































































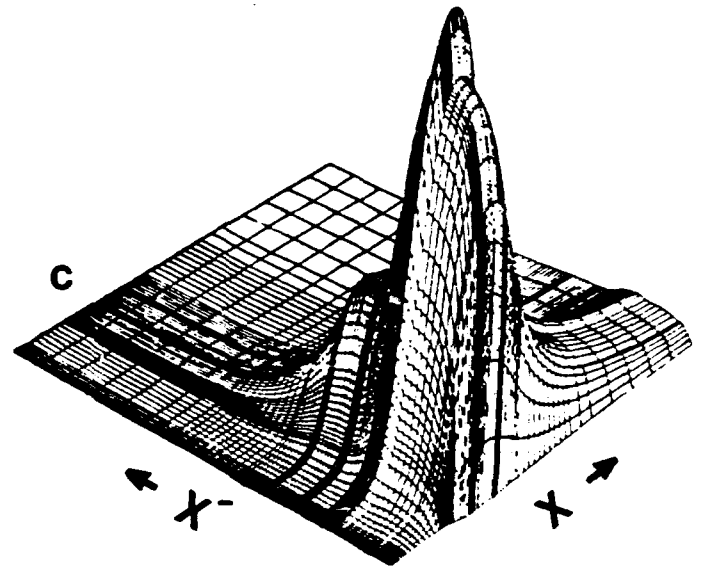
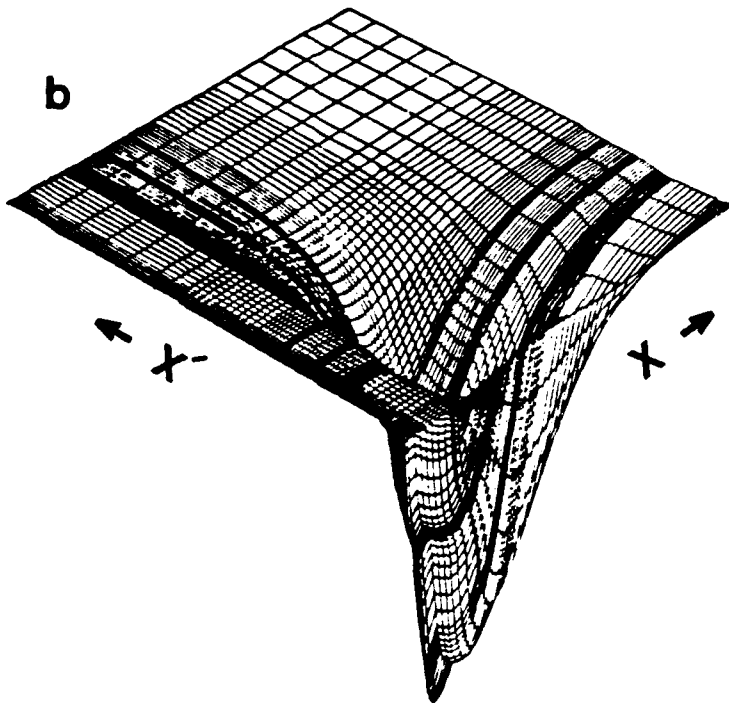
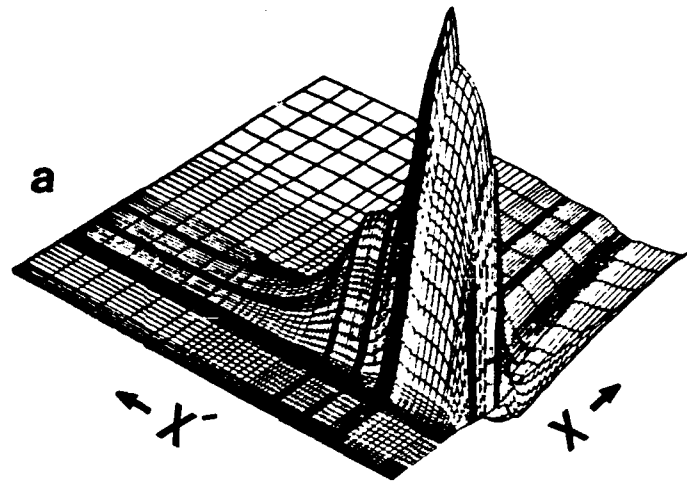
$t \backslash t'$	CO ₂	CO	O ₂	H ₂ O	H ₂	HCO	H	HO ₂	H ₂ O ₂	O	OH
CO ₂											
CO											
O ₂											
H ₂ O											
H ₂											
HCO											
H											
HO ₂											
H ₂ O ₂											
O											
OH											

Figure 18

$x' \backslash x$	CO2	CO	O2	H2O	H2	HCO	H	HO2	H2O2	O	OH
CO2											
CO											
O2											
H2O											
H2											
HCO											
H											
HO2											
H2O2											
O											
OH											

Figure 19



Appendix E

5. Sensitivity Analysis of a Steady-state Premixed Laminar $\text{CO}+\text{H}_2+\text{O}_2$ Flame, M. Mishra, R. Yetter, Y. Reuven, and H. Rabitz, Int. J. Chem. Kinetics, submitted.

Sensitivity Analysis of a Steady-state, Premixed
Laminar CO-H₂-O₂ Flame

Manoj Mishra
Department of Chemistry
Indian Institute of Technology
Powai, Bombay, 400076 India

Richard Yetter
Yakir Reuven and Herschel Rabitz
Department of Chemistry
Princeton University
Princeton, New Jersey 08540

and

Mitchell D. Smooke
Department of Mechanical Engineering
Yale University
New Haven, Connecticut

Submitted to Int. J. of Chem. Kinetics, 2/91

ABSTRACT

The direct and very efficient Newton method for obtaining sensitivities of two-point boundary value problems is utilized for detailed exploration of a reacting-diffusing $\text{CO}+\text{H}_2+\text{O}_2$ steady-state premixed laminar flame. Sensitivity coefficients and Green's functions calculated for this system offer exhaustive characterization and new insights into the role of diffusion and exothermicity in carbon monoxide oxidation kinetics. In particular, the reactions of the hydroperoxy radical with hydrogen, oxygen and hydroxyl radicals are found to be extremely important at all temperatures in the fuel lean (40 torr) flame studied here. The diffusive mixing of chemical species from the low and the high temperature portions of the flame and the large heats of reaction associated with the hydroperoxy radicals are found to be responsible for the increased importance of these reactions.

I. INTRODUCTION

The wet oxidation of carbon monoxide involves elementary steps common to the high temperature flame oxidation of all hydrocarbons. As such, the quest for a comprehensive mechanism for this system is an ongoing concern of fundamental importance to combustion chemistry and has been studied extensively in combustion kinetics¹. Recently, Yetter et al² have put forth a comprehensive mechanism for this system and have examined its validity over a wide range of experimental conditions in the absence of mass and energy transfer. In addition, they performed a thorough sensitivity analysis of its temporal kinetics³. In another paper⁴, this same system was studied using a one-step "global" reaction. In this latter work, the overall reaction was represented by the single step, $\text{CO} + 1/2 \text{O}_2 \rightarrow \text{CO}_2$, with the reaction rate defined as $d[\text{CO}]/dt = -k_{\text{OV}}[\text{CO}][\text{H}_2\text{O}]^{1/2}[\text{O}_2]^{1/4}$. From results in which the overall rate constant, k_{OV} , was deduced from detailed calculations using the above elementary reaction mechanism, it was observed that the behavior of k_{OV} as a function of temperature was significantly different for premixed flames versus various temporal problems. Although one might anticipate that the wet carbon monoxide oxidation reaction may be strongly influenced by transport processes, no detailed study exists which explicitly demonstrates and explains the interplay between chemical kinetics and diffusion phenomena. Such an investigation is the principal concern of the present paper.

Validation of a reaction mechanism involves a detailed analysis of the effect of the changes in underlying input parameters (e.g., reaction rate constants, reactant flow rates, diffusion coefficients, etc.) on the experimental outputs (e.g., the concentration profiles). A systematic probe of the relationship between the output information obtained from a model and the input parameters (including the initial and boundary values) defining the

model constitutes the basic domain of sensitivity analysis. In recent years, sensitivity analysis has emerged as a potent tool for numerical investigation and validation of physico-mathematical models⁵. The major obstacle in the systematic calculation of sensitivity information has been the amount of additional computation required in solving the sensitivity equations which can easily exceed the computational effort required in obtaining the model results alone. This can be prohibitively expensive for models consisting of a large system of differential equations.

Recently, we have implemented a direct and very efficient approach for obtaining sensitivities of two-point boundary value problems using Newton's method⁶. Application of this procedure in the present paper to a reacting-diffusing $\text{CO}+\text{H}_2+\text{O}_2$ steady-state premixed laminar flame offers fresh insights regarding the role of diffusion in combustion chemistry. In Section II we present a brief description of the method for solving the differential equations governing the reacting-diffusing system in a steady-state flame and the calculations of the corresponding sensitivity coefficients. The species and their sensitivity profiles are analyzed in Sections III and IV, respectively. The premixed flame results are then compared to the results from pure temporal kinetics in Section V. In particular, diffusion and reaction exothermicity on the underlying kinetics is examined in detail, where we identify the conditions that offer a formal similarity between equations governing pure temporal kinetics and reacting-flowing steady-state kinetics. Finally, in Section VI concluding remarks summarize our major findings from this investigation.

II. Sensitivity Analysis of Reacting-Flowing Systems

The wet oxidation of carbon monoxide in a steady, one-dimensional, premixed laminar flame is modelled as a two-point boundary value problem. The formulation of the problem we consider closely follows the one originally proposed by Hirschfelder and Curtiss⁷. Upon neglecting viscous effects, body forces, radiative heat transfer and the diffusion of heat due to concentration gradients, the equations governing the structure of a steady one-dimensional isobaric flame are

$$\dot{M} = \rho u = \text{constant} \quad (2.1)$$

$$\dot{M} \frac{dY_k}{dx} = - \frac{d}{dx} (\rho Y_k V_k) + \dot{\omega}_k W_k, \quad k = 1, 2, \dots, K, \quad (2.2)$$

$$\dot{M} \frac{dT}{dx} = \frac{1}{c_p} \frac{d}{dx} \left(\lambda \frac{dT}{dx} \right) - \frac{1}{c_p} \sum_{k=1}^K \rho Y_k V_k c_{p_k} \frac{dT}{dx} - \frac{1}{c} \sum_{k=1}^K \dot{\omega}_k h_k W_k \quad (2.3)$$

$$\rho = \frac{p \bar{W}}{RT} \quad (2.4)$$

In these equations x denotes the independent spatial coordinate fixed to the flame; \dot{M} , the mass flow rate; T , the temperature; Y_k , the mass fraction of the k -th species; p , the pressure; u , the velocity of the fluid mixture; ρ , the mass density; W_k , the molecular weight of the k -th species; \bar{W} , the mean molecular weight of the mixture; R , the universal gas constant; λ , the thermal conductivity of the mixture; c_p , the constant pressure heat capacity of the mixture; c_{p_k} , the constant pressure heat capacity of the k -th species; $\dot{\omega}_k$, the molar rate of production of the k -th species per unit

volume; h_k , the specific enthalpy of the k -th species; and V_k , the diffusion velocity of the k -th species. The form of the chemical production rates and the diffusion velocities can be found in references 8 and 9.

The problem is posed on the infinite interval $-\infty < x < \infty$ with the boundary conditions at $x = -\infty$ given by

$$T(-\infty) = T_u, \quad (2.5)$$

$$Y_k(-\infty) = Y_{k_u}, \quad k = 1, 2, \dots, K, \quad (2.6)$$

and at $x = \infty$ by

$$\frac{dT}{dx}(\infty) = 0, \quad (2.7)$$

$$\frac{dY_k}{dx}(\infty) = 0, \quad k = 1, 2, \dots, K, \quad (2.8)$$

where the Y_{k_u} are the specified mass fractions of the reactants and T_u is the temperature of the unreacted gas. We point out that instead of solving the governing equations on the infinite domain, we pose the problem on the finite interval $0 \leq x \leq L$ where the length of the interval must be large enough to insure that the boundary conditions are properly satisfied¹⁰. The new boundary conditions at $x = 0$ are given by

$$T(0) = T_u, \quad (2.9)$$

$$\epsilon_k^{(0)} = Y_{k_u}, \quad k = 1, 2, \dots, K, \quad (2.10)$$

and at $x = L$ by

$$\frac{dT}{dx}(L) = 0, \quad (2.11)$$

$$\frac{dY_k}{dx}(L) = 0, \quad k = 1, 2, \dots, K, \quad (2.12)$$

where the mass flux of the k -th species is defined as

$$\epsilon_k = Y_k + \frac{\rho Y_k V_k}{\dot{M}}, \quad k = 1, 2, \dots, K. \quad (2.13)$$

We point out that in an adiabatic problem, the mass flow rate \dot{M} is not known; it is an eigenvalue to be determined. Calculation of the flow rate proceeds by introducing the trivial differential equation

$$\frac{d\dot{M}}{dx} = 0, \quad (2.14)$$

and an additional boundary condition to the system in (2.1-2.13). The particular choice of the extra boundary condition is somewhat arbitrary but, it must be chosen, however, to insure that the spatial gradients of both the temperature and the mass fractions are vanishingly small at $x=0$. In keeping with the dominant role of the temperature, we have chosen to fix the temperature at an interior grid point such that

$$T(x_f) = T_f, \quad (2.15)$$

where x_f is a specified spatial coordinate interior to the domain and T_f is a specified temperature. Values of x_f and T_f should be chosen to guarantee a nearly zero temperature gradient at the unreacted boundary.

Solution of the governing equations proceeds with an adaptive nonlinear boundary value method on an initial mesh containing m grid points. Upon discretization of the differential operators in (2.1-2.15), we obtain a system of nonlinear algebraic equations

$$F(U, \underline{\alpha}) = 0, \quad (2.16)$$

where U represents the vector of N dependent variables, and the vector $\underline{\alpha}$ of length M represents the system parameters such as activation energies, pre-exponential factors and other quantities that enter the differential equations. Solution of the system in (2.16) by Newton's method has been discussed in detail elsewhere and we refer the reader to the appropriate references (see, e.g., references 10 and 11).

In keeping with our goal of ascertaining the role and importance of various system parameters, the quantities of natural interest are the first-order sensitivity coefficients

$$s_{ij} = \frac{\partial U_i(x, \underline{\alpha})}{\partial \alpha_j}, \quad (2.17)$$

which provide a direct measure of how the j -th parameter controls the behavior of the i -th dependent variable at point x . The appropriate equations for these quantities can be derived by differentiating (2.16) with respect to α_j . We have

$$\frac{d}{d\alpha_j} (F(U, \underline{\alpha})) = \frac{\partial F}{\partial U} \frac{\partial U}{\partial \alpha_j} + \frac{\partial F}{\partial \alpha_j} = 0, \quad j = 1, 2, \dots, M. \quad (2.18)$$

Recalling that the Jacobian matrix is given by $J = \partial F / \partial U$, we have

$$J \frac{\partial U}{\partial \alpha_j} = - \frac{\partial F}{\partial \alpha_j}, \quad j = 1, 2, \dots, M. \quad (2.19)$$

Although equation (2.18) can be solved at any level of the Newton iteration and at any level of grid refinement, we solve it on the finest grid with the last Jacobian formed. It is only at this stage of the calculation that the numerical solution has been resolved with sufficient accuracy to represent the true solution.

We point out that although the original boundary value problem is nonlinear, the sensitivity equations in (2.19) are linear. In principle, we can apply the Green's function method to obtain a solution to (2.19). While we do not advocate such a procedure, the Green's function does, however, contain valuable information on system sensitivity. The Green's function satisfies the equation

$$JG = -\Delta \quad (2.20)$$

where the diagonal matrix Δ can be written in terms of $N \times N$ diagonal blocks δ_j , $j = 1, 2, \dots, m$. To insure that the Green's function vanishes at the boundaries, the diagonal blocks corresponding to $j = 1$ and $j = m$ are set identically to zero. The nonzero diagonal entries of the remaining blocks $j = 2, 3, \dots, m-1$ are given by

$$(\delta_j)_{kk} = \frac{2}{h_j + h_{j+1}}, \quad k = 1, 2, \dots, N, \quad (2.21)$$

where h_j , $j = 2, 3, \dots, m$ is the j -th mesh interval. With the definition in (2.21) we can obtain G by solving the linear system $JG = -\Delta$. Assuming the Jacobian has been factored, formation of G is accomplished by performing Nm back substitutions with a different column of Δ as the right-hand side.

The elements of G have the response function interpretation¹²

$$G_{ij}(x, x') = \frac{\delta Y_i(x)}{\delta J_j(x')} \quad (2.22)$$

i.e., the elements $G_{ij}(x, x')$ correspond to the response of the i -th dependent variable at point x to a disturbance of the flux $J_j(x')$ of the dependent variable j at point x' . The solution to Eq. (2.19) may now be expressed in terms of the Green's function

$$S_{ij}(x) = \sum_{\ell} \int_0^L dx' G_{i\ell}(x, x') g_{\ell j}(x'). \quad (2.23)$$

The fundamental role of the Green's function is self-evident from its interpretation in Eq. (2.22) and its role in Eq. (2.23). In particular, from Eq. (2.23) all the system sensitivities are expressed in terms of a convolution of the Green's function with the explicit parametric derivatives of the differential equations.

A detailed account of the numerical procedure and error analysis for direct calculation of the system sensitivities and Green's functions using an adaptive finite difference technique with Newton's method has been discussed

elsewhere⁶. This is the method we have used for obtaining the sensitivity coefficients and Green's functions for the $\text{CO}+\text{H}_2+\text{O}_2$ system. The physical content and significance of this latter information is analyzed in the following sections.

III. The $\text{CO}+\text{H}_2+\text{O}_2$ System

The present analysis is performed on a laminar, premixed, fuel-lean, $\text{CO}+\text{H}_2+\text{O}_2$ flame. This particular flame has been experimentally studied using a 5 cm cylindrical burner by Vandooren, Peters, and van Tiggelen¹³ and it has been modelled numerically by Cherian et al¹⁴. The composition of the unburnt gas (i.e., the upstream conditions) in mole fractions was $X_{\text{CO}} = 0.094$, $X_{\text{H}_2} = 0.114$, and $X_{\text{O}_2} = 0.792$. The temperature and pressure of the unburnt gas were 273 K and 40 Torr, respectively.

The calculated adiabatic temperature and species mole fraction profiles of the reactants, intermediates, and products are shown in Figure 1. These calculations were based on a comprehensive reaction mechanism^{2,3} consisting of 27 reversible reactions and 11 chemical species (see Table 1). The mechanism differs from the previous numerical work mainly in the presence of H_2O_2 and its associated reactions. The dynamic role of hydrogen peroxide will be discussed later along with the analysis of the sensitivity gradients.

The results for the species concentrations are in good agreement with the experimental data and the earlier numerical results although the present flame speed is approximately 25% larger than the experimental value. As discussed by Cherian et al¹⁴, the experimental data were taken under conditions of non-negligible heat losses to the burner surface. These energy losses were not incorporated into the present calculations (due to lack of

experimental information on the rate of heat abstraction and the temperature of the burner) nor were chemical losses included such as catalytic recombination of radicals at the burner surface. Moreover, the sensitivity analysis results presented in the next section show that small uncertainties in many of the model input parameters may contribute to the flame speed difference. The present model is run under well-defined conditions that allow us to investigate the role of the various physical parameters on the structure of a $\text{CO}+\text{H}_2+\text{O}_2$ premixed flame, and, in particular, to study the effects of molecular diffusion and temperature on the controlling chemistry.

IV. Analysis of Linear Sensitivity Gradients

The normalized linear gradients of the CO concentration profile with respect to various reaction rate constants and diffusion coefficients are shown in Figure 2. The sensitivity of CO with respect to the system pressure, the mixture thermal conductivity and the total mass flow rate are shown in Figure 3. From these figures, a ranking of the relative importance of these variables on the CO concentration may be obtained. The importance of the variables was also found to be the same with regard to the flame speed but naturally of opposite sign (see Table 2) due to the inverse relation of the flame speed and the CO concentration under the present running conditions. These flame speed sensitivities were obtained from a "derived" sensitivity analysis⁶.

Underlying microscopic processes to which the CO concentration profile and also the flame speed are most sensitive are the elementary reaction of CO with the hydroxyl radical (i.e., reaction 11 in table 1 which is the rate controlling step of the overall reaction rate, RR) and to the mixture thermal conductivity (the most sensitive parameter of the molecular transport

processes). The results agree with traditional phenomenological analyses¹⁵ which yield

$$u \sim \sqrt{\Lambda(RR)} \quad (4.1)$$

where $\Lambda = (\lambda/\rho C_p)$ is the thermal diffusivity of the mixture. This relationship is manifested in that

$$\frac{\partial \ln u}{\partial \ln \lambda} \approx \frac{\partial \ln u}{\partial \ln k_{11}} \quad (4.2)$$

(as also seen in the CO sensitivity profiles of Figs. (2a) and (3)). The sensitivity gradients of Figure 2 also show the inhibiting effects of H₂ and CO diffusion, and the accelerating effects of H₂O and OH diffusion. This behavior is also consistent with the role played by the species in the flame. However, it is interesting to note that H₂ diffusion is nearly as important as H-atom diffusion and that CO diffusion is as important as OH-radical diffusion.

The relative importance of other reactions is also easily recognized from Figure 2. It may be seen that the branching reaction $H + O_2$ (i.e., rate constants 15 and 16) is nearly microscopically balanced which is reflected in the relation $\partial \ln CO / \partial \ln k_{15} \approx - \partial \ln CO / \partial \ln k_{16}$. The net sensitivity of $H_2 + O = H + OH$ is in the forward direction and the net sensitivity of $O + H_2O = 2OH$ is in the reverse. The two propagating reactions $CO + OH \rightarrow H + CO_2$ and $H_2 + OH \rightarrow H + H_2O$ each promote the overall reaction while the recombination steps $H + O_2 + M \rightarrow HO_2 + M$, $H + OH + M \rightarrow H_2O + M$ inhibit the reaction. These results parallel findings on the same purely temporal analogous problem^{2,3} with one exception: in the temporal problems studied, $H_2 + OH \rightarrow H + H_2O$ was generally found to

inhibit (or slow) the CO reaction, whereas here it is seen to promote the reaction.

An interesting outcome from the present set of calculations is the evident relative importance of the radical-radical reactions involving HO₂ (e.g., H + HO₂, OH + HO₂). For example, observe that the CO concentration is more sensitive to the reaction H + HO₂ → OH + OH than to H + O₂ → OH + O at all temperatures (i.e., even in the post-flame region). One explanation for this occurrence results from the near equilibration of H + O₂ → OH + O. As a consequence, few H-atoms are removed from the system by this reaction. This makes the termolecular reaction H + O₂ + M → HO₂ + M (which like other recombination reactions has a rate with a negative temperature dependence) competitive for H-atoms at high temperatures and therefore allows for further HO₂ reactions. These reactions receive further attention in the next section.

Also noticeable in the sensitivity gradients is a remarkable similarity between various profiles irrespective of the parameter being perturbed, except in the case of $\partial \ln \text{CO} / \partial \ln k_{12}$ where there is a loss of similarity in the post-flame region (see Figure 2c). When strong coupling exists between several dependent variables and a single variable (such as the temperature) dominates the behavior of the others, the sensitivity gradients may be scaled¹⁶ in the following fashion

$$\left[\frac{\partial Y_n(x)}{\partial \alpha_j} \right] \approx \left[\frac{\partial T(x)}{\partial \alpha_j} \right] \left[\frac{\partial Y_n}{\partial x} \right] \left[\frac{\partial T}{\partial x} \right]^{-1} \quad (4.3)$$

For example, the gradients for the CO concentration all pass through zero at a position of $x = 0.55$ cm. The change in sign of the sensitivity gradients

here is directly correlated to the change in sign of the CO slope, dCO/dx , at that point. The (slight) positive growth in the CO concentration prior to significant reaction results from preferential diffusion of the lighter molecular and atomic weight species. This is evident in Figure 4 which shows the atomic hydrogen to atomic carbon ratio (H/C) as a function of flame position. As can be seen from the figure, the overall mixture is deficient of hydrogen containing species where the CO concentration peaks and slightly thereafter. However, more importantly this relationship (Eq. 4.3) aids the analysis of the $CO + H_2 + O_2$ flame in that the ranking of reactions for one dependent variable profile are sufficient to determine the overall importance of reactions in the entire mechanism.

An increase in pressure is observed to decrease significantly the CO concentration (Figure 3) or equivalently to increase the flame speed. Increasing the total mass flow rate decelerates the reaction and eventually leads to unstable conditions.

The role of hydrogen peroxide in the flame structure is illustrated by considering the effects of its elementary steps on the other species of the system. Figure 5a shows such gradients of the CO concentration and Figure 5b shows the gradients of the O-atom concentration. The major steps producing hydrogen peroxide are hydroxyl radical recombination, $OH + OH + M = H_2O_2 + M$, and $HO_2 + HO_2 = H_2O_2 + H_2$. Clearly, the CO concentration is not altered noticeably, but significant changes are observed in the O-atom concentration (as well as the other intermediates) in the low temperature regime of the flame. Even throughout the rest of the flame, the presence of H_2O_2 and its associated reactions have some influence. At higher pressure, H_2O_2 would be expected to play a more important role due to the increase in HO_2 concentrations as a result of the reaction $H + O_2 + M \rightarrow HO_2 + M$ dominating H

+ O₂ → OH + O and hence, its inclusion is recommended for all comprehensive mechanisms.

V. Comparison of Flame Chemistry with Dilute Temporal Chemistry

Comparison of the dominant elementary steps described in the last section with those steps generally found important in temporal systems² reveals some dramatic differences in the underlying mechanism of the two systems. Using the same reaction mechanism, shock tube and flow reactor data were modelled in a previous paper² and through a similar sensitivity analysis, the controlling reactions on the CO concentration were investigated. In these latter models, diffusive processes are assumed small compared to the remaining terms in Eq. (2.2). In addition, the chemistry is performed under nearly isothermal conditions. The simplest mathematical equations governing these two experimental systems are

$$\rho \frac{d\bar{Y}_k}{dt} = \dot{\omega}_k W_k \quad (5.1)$$

and

$$\dot{M} \frac{dY_k}{dx} = \dot{\omega}_k W_k \quad (5.2)$$

for the shock tube and flow reactor, respectively. Here, $\dot{\omega}_k$ is the identical reaction matrix found in Eq. (2.2). When u is constant, the analysis of both systems is identical since $u dY_k/dx$ may be equated to dY_k/dt .

One difference with the pure temporal chemistry mentioned earlier manifests itself in the role of radical-radical reactions of HO₂ with H, O, and OH. In the temporal systems, these reactions were usually of secondary importance and never exceeded the importance of the principal reactions of

the hydrogen-oxygen mechanism such as $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$, $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$, $\text{O} + \text{H}_2\text{O} \rightarrow \text{OH} + \text{OH}$, and $\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$.

The point to be emphasized here is that a change in the important steps of the reaction mechanism is apparent between temporal and flame problems. In the two sections to follow, two specific aspects of the flame problem not encountered in the temporal problems are considered as responsible for these significant differences. First, the role of diffusion and second, the role of mixture exothermicity on the chemistry are investigated.

Va. Diffusion Effects

The role of diffusion can be examined from several different perspectives. From the system Green's function, it is evident that significant diffusion is present which may affect the chemistry. This is illustrated in Figure 6 which shows the response surface corresponding to the Green's function matrix element $\delta\text{CO}_2(x)/\delta J_{\text{H}_2}(x')$. This figure reveals the response of the CO_2 concentration at position x in the profile to a disturbance of the H_2 species flux at position x' . Although diffusion occurs throughout the flame, the effect is clearly seen and separated from convective transport in the regions of $x' > x$. In particular in the latter upstream region a non-zero response of CO_2 can be ascribed to diffusion. The consequences of this effect on the sensitivity spectrum and the attendant chemistry have been noted in the previous section and contrasted with the results from the pure temporal problem².

The role of diffusion in the flame may also be investigated by examining a flow reactor system under mass flow conditions which diminish the role of the diffusion terms and thereby simulate the pure temporal problem. We

illustrate this by examining the analytic results from a simple linear kinetics problem modelled by

$$D \frac{d^2 Q}{dx^2} - \dot{M} \frac{dQ}{dx} + \underline{K} Q = \underline{Q}, \quad (5.3)$$

which may be equivalently written as

$$D \frac{d^2 Y}{dx^2} - \dot{M} \frac{dY}{dx} + \underline{\lambda} Y = \underline{Q}, \quad (5.4)$$

where $\underline{\lambda} = \underline{U}^{-1} \underline{K} \underline{U}$ is the diagonal matrix of the eigenvalues of \underline{K} and $\underline{Y} = \underline{U}^{-1} \underline{Q}$ and D is a chemistry weighted diffusion coefficient. The general solution to Eq. (5.4) is a linear combination

$$Y_i(x) = \sum_n \left[a_{in} \phi_n^1(x) + b_{in} \phi_n^2(x) \right] \quad (5.5)$$

where

$$\begin{aligned} \phi_n^1(x) &= \exp \left[\left[\frac{\dot{M}}{2D} + \sqrt{\frac{\dot{M}^2 - 4D\lambda_n}{2D}} \right] x \right], \\ \phi_n^2(x) &= \exp \left[\left[\frac{\dot{M}}{2D} - \sqrt{\frac{\dot{M}^2 - 4D\lambda_n}{2D}} \right] x \right], \end{aligned} \quad (5.6)$$

and $i = 1, \dots, K$ and $n = 1, \dots, K$.

The Green's function for Eq. (5.3) may be expressed in terms of the diagonal Green's function $G_{nn}(x, x')$ for Eq. (5.4). The latter function satisfies the boundary conditions $G_{nn}(\infty, x')|_{x \rightarrow \pm\infty} = 0$ and it can be constructed from the linearly independent solutions $\phi_n^1(x)$ and $\phi_n^2(x)$. For $\dot{M} \gg 4D|\lambda_n|$, these are easily shown to be

$$G_{nn}(x, x') = \frac{2\dot{D}\dot{M}}{(\dot{M}^2 - 2\lambda_n D)} \exp\left[\frac{\lambda_n}{\dot{M}}(x - x')\right], \quad x > x' \quad (5.7a)$$

$$\frac{2\dot{D}\dot{M}}{(\dot{M}^2 - 2\lambda_n D)} \exp\left[\frac{\dot{M}}{D}(x - x')\right] \exp\left[-\frac{\lambda_n}{\dot{M}}(x - x')\right], \quad x' > x \quad (5.7b)$$

Physically stable solutions exist for the eigenvalues being negative semidefinite $\lambda_n \leq 0$ and with this observation we can simply analyze the Green's function elements in Eq. (5.7). In the downstream region $x > x'$ exponential decay occurs from the point of disturbance x' dictated by λ_n/\dot{M} . This behavior is exactly reminiscent of what is found in purely temporal kinetics where the variables t and t' have an analogous meaning to x and x' . In addition, a temporal kinetics system would also have the Green's function being strictly zero for $t' > t$ due to causality and the analogous region in the present reaction-diffusion-convection problem is $x' > x$. From Eq. (5.7b) it is evident that the Green's function in the present problem is not strictly zero in this regime with the parameter \dot{M}/D playing a critical role. In particular, for larger values of \dot{M}/D the Green's function will decay more rapidly from the point of disturbance x' in the upstream regime $x' > x$. This behavior is physically reasonable and can be viewed as arising due to a diminution of the diffusion coefficient D .

The Green's function matrix results discussed above are consistent with the linear parametric gradients of Section IV. The CO mole fraction was observed to be highly sensitive, with the same directional sense, as the total mass flow rate and the diffusive coefficients upon the reactants H_2 and CO. Here, increasing the diffusion coefficients of H_2 and CO adds to the overall mass flux into the flame front decreasing the effectiveness of

radical transport upstream. In the context of the present analysis (i.e., examining the differences in kinetics between transport-free systems and flames), most of these differences, particularly the changes in importance of the bi-molecular radical-radical reactions (as discussed above), are explainable through the effect of transport on the mixing of the low (HO_2 , H_2O_2) and the high (H , O , OH) temperature intermediate species.

Vb. Effect of Mixture Exothermicity

Another significant difference between the present flame problem and the analogous temporal system is the overall exothermicity of the mixtures studied. In the shock tube and the flow reactor experiments, the mixtures were all dilute and thus nearly thermo-neutral. The flame problem on the other hand is extremely exothermic, and much of the controlling chemistry in the flame can be attributed to the rapid temperature rise. The heats of reaction of several of the elementary steps, found to be important in the flame problem, are listed in Table 3. It is apparent that the HO_2 production and consumption reactions are all very exothermic.

The degree to which the temperature plays an intermediary role in making a particular reaction important can be investigated using "reduced" Green's function techniques¹⁷. In this technique, the response of the temperature may be frozen at its nominal spatial dependence $T(x)$ and shielded from other perturbations introduced to the system. Hence, the dynamic couplings between the chemical species may be examined without temperature responses playing a role. We should emphasize that this temperature constrained calculation does not effect the structure of the flame in any way; only the sensitivities will differ. Some linear sensitivity gradients obtained with the frozen temperature profile are shown in Figure 7. The importance of the HO_2 -

radical reactions is greatly reduced and much of the self-similarity in the gradient profiles has disappeared. This disappearance of the self-similarity confirms the dominant controlling role of the temperature¹⁶ which enters the problem exponentially whereas all other dependent variables (i.e., species) enter linearly or quadratically.

Finally, it is also interesting to note that of all the reactions, only one coefficient has changed directional sense, i.e., $\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$. In the constrained temperature problem, this reaction inhibits CO oxidation, much as was found in previous temporal problems, since the heat release from this reaction is now not available to accelerate the overall reaction as was found in the original flame problem. Hence, the reaction exothermicity not only can change which reactions are important, but also the role these reactions play.

VI. Concluding Remarks

In the present paper, modelling and sensitivity analysis techniques have been applied to study the structure of a premixed $\text{CO} + \text{H}_2 + \text{O}_2$ flame. Our analysis has shown that the presence of molecular transport alters the chemistry of this system. Furthermore, the exothermicity of the mixture also affects the chemistry. Both of these results are particularly important with regard to the development, application, and validation of reaction mechanisms. Specifically, the fast reactions of HO_2 with H , OH , and O were found to be important at all positions throughout the low-pressure, lean flame studied here. Accurate rate data for these reactions at temperatures above 1000 K are therefore of obvious importance. Also, although the presence of hydrogen peroxide in the mechanism is found to have little influence on major species, temperature and the flame speed, we find it to be

of considerable importance with regard to the concentration of other intermediates.

Lastly, some general comments on chemical kinetic studies in flames, flow reactors, and shock tubes may be reasoned from the present results. Without a doubt, the simplest of these experiments to interpret kinetics data are from shock tubes and flow reactors. This can readily be seen from the differential equations governing these systems. However, because of their practical importance, the kinetics of flames must continue to be studied particularly since the dominant reaction pathways may differ from those found in simpler transport-free experiments. Furthermore, data from premixed flames are necessary to validate heat release rates and flame speed predictions. In premixed flames, the kinetics and transport are of almost equal importance; however, the system is almost entirely driven by the heat release and hence the temperature profile through the flame. The ability to deconvolute the kinetics, which produce this heat release, is very difficult due to simultaneous transport processes and the high sensitivity of measured observables to the temperature measurements. Hence parameter extraction/verification from flame studies is more difficult than in shock tubes or flow reactors, as inferred from the present $\text{CO} + \text{H}_2 + \text{O}_2$ flame by the high degree of coupling among parameters and consequently, such evaluations should generally be carried out in the simpler systems.

Acknowledgment

We acknowledge the support for this work from the Department of Energy and the Air Force Office of Scientific Research.

References

1. C.K. Westbrook and F.L. Dryer, Prog. Energy Combust. Sci., 10, 1 (1984).
2. R.A. Yetter, F.L. Dryer and H. Rabitz, "A Comprehensive Reaction Mechanism for Carbon Monoxide-Hydrogen-Oxygen Kinetics", Western States Section, The Combustion Institute, Fall Technical Meeting, Paper No. W5S84-96, Stanford University, Palo Alto, October 1984. Also R.A. Yetter, F.L. Dryer and H. Rabitz, Combust. Sci. Tech., in press 1990, report a revised version of this mechanism with more recent rate constant evaluations. The newer mechanism produces similar results with a flame speed approximately 12% greater than the experimental value.
3. R.A. Yetter, F.L. Dryer and H. Rabitz, Combustion and Flame, 59, 107 (1985).
4. R.A. Yetter, F.L. Dryer and H. Rabitz, Twenty-first Symposium (International on Combustion, The Combustion Institute, Pittsburgh, PA 1986.
5. H. Rabitz, M. Kramer, and D. Dacol, Ann. Rev. Phys. Chem., 34, 419 (1983); H. Rabitz, Computers and Chemistry, 5, 167 (1981).
6. Y. Reuven, M.D. Smooke and H. Rabitz, J. Comp. Phys., 64, 27 (1986).
7. J.O. Hirschfelder and C.F. Curtiss, J. Chem. Phys., 17, 1076 (1949).
8. R.J. Kee, J.A. Miller, and T.H. Jefferson, Sandia National Laboratories Report SAND80-8003 (1980); R.J. Kee, J. Warnatz and J.A. Miller, "A Fortran Computer Code Package for the Evaluation of Gas-phase Viscosities, Conductivities and Diffusion Coefficients", Sandia National Laboratories Report, SAND83-8209, (1983).

9. J. T. Hwang, E.P. Dougherty, S. Rabitz and H. Rabitz, J. Chem. Phys., 69, 5180 (1978); E.P. Dougherty, J.T. Hwang and H. Rabitz, J. Chem. Phys., 71, 1794 (1979).
10. M.D. Smooke, J.A. Miller and R.J. Kee, Combust. Sci. Tech., 34, 79 (1983).
11. M.D. Smooke, J. Comp. Phys., 48, 72 (1982).
12. M. Demiralp and H. Rabitz, J. Chem. Phys., 74, 3362 (1981); *ibid*, 75, 1810 (1981).
13. J. Vandooren, J. Peters and P.J. van Tiggelen, Fifteenth Symposium (International) on Combustion, p. 745, The Combustion Institute, 1975.
14. M.A. Cherian, P. Rhodes, R.J. Simpson and G. Dixon-Lewis, Eighteenth Symposium (International) on Combustion, The Combustion Institute, 1981.
15. I. Glassman, Combustion, (Academic Press, New York, 1977).
16. H. Rabitz and M.D. Smooke, J. Phys. Chem., 92, 1110 (1988).
17. M. Mishra, L. Peiperl, Y. Reuven, H. Rabitz and M.D. Smooke, "On the Use of Green's Functions for the Analysis of Dynamic Couplings: Some Examples from Chemical Kinetics and Quantum Dynamics", in press.

Figure Captions

- Figure 1. Species and Temperature profiles for the sample flame.
- Figure 2. Normalized sensitivities of the CO mole fraction profile with respect to various reaction rate constants. In Figure 2(a-c) the numbers labelling the various curves correspond to the elementary steps from Table I. In figure 2d, D_X denotes the diffusion coefficient of species X.
- Figure 3. Normalized sensitivities of the CO mole fraction profile with respect to the system mass flow rate \dot{M} , pressure P and the thermal conductivity λ .
- Figure 4. Ratio of total Hydrogen to total Carbon in the fuel mixture.
- Figure 5a. Sensitivity gradients of the CO mole fraction profile with respect to rate constants of various H_2O_2 reaction. Conventions of Figure 1 apply.
- Figure 5b. Sensitivity gradients of the O-atom mole fraction profile with respect to rate constants of various H_2O_2 reactions. Conventions of Figure 1 apply.
- Figure 6. Green's function surface $\delta CO_2(x)/\delta J_{H_2}(x')$ corresponding to the response of CO_2 to a perturbation in the flux of H_2 .
- Figure 7. Sensitivities of the CO mole fraction profile to various reaction rates for the original flame but with a frozen temperature profile. Conventions of Figure 1 apply.

TABLE 1. CO+H₂+O₂ Kinetic Mechanism

INDEX	REACTION	A ¹	n	E	I ²
1,2 ³	HCO + H = CO + H ₂	2.00(14) ⁴	0.0	0.0	f
3,4	HCO + OH = CO + H ₂ O	1.00(14)	0.0	0.0	f
5,6	O + HCO = CO + OH	3.02(13)	0.0	0.0	f
7,8	HCO + O ₂ = CO + HO ₂	3.01(12)	0.0	0.0	f
9,10	CO + HO ₂ = CO ₂ + OH	1.50(14)	0.0	2.36(4)	f
11,12	CO + OH = H + CO ₂	4.46(6)	1.5	-7.40(2)	f
13,14	CO ₂ + O = CO + O ₂	2.53(12)	0.0	4.77(4)	b
15,16	H + O ₂ = O + OH	3.73(17)	-1.0	1.75(4)	f
17,18	H ₂ + O = H + OH	1.80(10)	1.0	8.90(3)	f
19,20	O + H ₂ O = OH + OH	4.58(9)	1.3	1.71(4)	f
21,22	H + H ₂ O = OH + H ₂	1.08(9)	1.3	3.65(3)	b
23,24	H ₂ O ₂ + OH = H ₂ O + HO ₂	7.00(12)	0.0	1.43(3)	f
25,26	HO ₂ + O = O ₂ + OH	1.81(13)	0.0	-3.97(2)	f
27,28	H + HO ₂ = OH + OH	1.69(14)	0.0	8.74(2)	f
29,30	H + HO ₂ = H ₂ O + O ₂	6.63(13)	0.0	2.13(3)	f
31,32	OH + HO ₂ = H ₂ + O ₂	1.45(16)	-1.0	0.0	f
33,34	H ₂ O ₂ + O ₂ = HO ₂ + HO ₂	1.00(13)	0.0	1.00(3)	b
35,36	HO ₂ + H ₂ = H ₂ O ₂ + H	1.70(12)	0.0	3.75(3)	b
37,38	O ₂ + M = O + O + M	6.17(15)	-0.5	0.0	b
39,40	H ₂ + M = H + H + M	2.20(14)	0.0	9.60(4)	f
41,42	OH + M = O + H + M	1.00(16)	0.0	0.0	b
43,44	H ₂ O ₂ + M = OH + OH + M	120(17)	0.0	4.55(4)	f
45,46	H ₂ O + M = H + O ₂ + M	2.20(16)	0.0	1.05(5)	f
47,48	HO ₂ + M = H + O ₂ + M	1.65(15)	0.0	-1.00(3)	b

- 27 -

INDEX	REACTION	A ¹	n	E	I ²
49,50	CO ₂ + M = CO + O + M	5.90(15)	0.0	4.10(3)	b
51,52	HCO + M = H + CO + M	6.90(14)	0.0	7.00(3)	b
53,54	H + H ₂ O ₂ = H ₂ O + OH	1.00(13)	0.0	3.59(3)	f

[M] = [N_w] + [O₂] + 16[H₂O] + 2.5[H₂] + 3.8[CO₂] + 1.9[CO] + [HO₂] + H₂O₂] + [H] + [O] + [OH] + [HCO] + 0.87[Ar]

¹ Units are cm-mole-sec-cal, $k = AT^n \exp(-E/RT)$

² I indicates direction of the reaction for which rate constant data are used. References for the rate data may be found in Reference 3.

³ Index associated with forward rate constant, reverse rate constant.

⁴ In this and all subsequent tables, numbers in parentheses denote powers of ten.

TABLE 2. Linear sensitivities of the flame speed
with respect to various pre-exponential factors

j	REACTION	$\partial \ln(\text{Flame Speed}) / \partial \ln A_j$
11	$\text{CO} + \text{OH} \rightarrow \text{CO}_2 + \text{H}$	42.1
12	$\text{CO}_2 + \text{H} \rightarrow \text{CO} + \text{OH}$	-0.8
15	$\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$	9.9
16	$\text{OH} + \text{O} \rightarrow \text{H} + \text{O}_2$	-9.2
17	$\text{H}_2 + \text{O} \rightarrow \text{OH} + \text{H}$	22.1
18	$\text{OH} + \text{H} \rightarrow \text{H}_2 + \text{O}$	-2.6
19	$\text{O} + \text{H}_2\text{O} \rightarrow \text{OH} + \text{OH}$	6.3
20	$\text{OH} + \text{OH} \rightarrow \text{O} + \text{H}_2\text{O}$	-13.4
22	$\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$	11.6
25	$\text{O} + \text{HO}_2 \rightarrow \text{OH} + \text{O}_2$	0.4
27	$\text{H} + \text{HO}_2 \rightarrow \text{OH} + \text{OH}$	13.2
29	$\text{H} + \text{HO}_2 \rightarrow \text{H}_2 + \text{O}_2$	-7.8
31	$\text{OH} + \text{HO}_2 \rightarrow \text{H}_2\text{O} + \text{O}_2$	-5.4
46	$\text{H} + \text{OH} + \text{M} \rightarrow \text{H}_2\text{O} + \text{M}$	-5.7
48	$\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$	-4.3

* sensitivities are evaluated at $x = 0.75$ cm

TABLE 3. Heats of reaction evaluated at 298 K

j	REACTION	$\Delta H_{298}(\text{kcal/mole})$
11	$\text{CO} + \text{OH} \rightarrow \text{CO}_2 + \text{H}$	-24.97
15	$\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{O}$	16.89
17	$\text{H}_2 + \text{O} \rightarrow \text{OH} + \text{H}$	1.97
20	$\text{OH} + \text{OH} \rightarrow \text{O} + \text{H}_2\text{O}$	-17.11
22	$\text{H}_2 + \text{OH} \rightarrow \text{H}_2\text{O} + \text{H}$	-15.13
25	$\text{O} + \text{HO}_2 \rightarrow \text{OH} + \text{O}_2$	-55.12
27	$\text{H} + \text{HO}_2 \rightarrow \text{OH} + \text{OH}$	-38.23
29	$\text{H} + \text{HO}_2 \rightarrow \text{H}_2 + \text{O}_2$	-57.10
31	$\text{OH} + \text{HO}_2 \rightarrow \text{H}_2\text{O} + \text{O}_2$	-72.23
40	$\text{H} + \text{H} + \text{M} \rightarrow \text{H}_2 + \text{M}$	-104.19
46	$\text{H} + \text{OH} + \text{M} \rightarrow \text{H}_2\text{O} + \text{M}$	-119.32
48	$\text{H} + \text{O}_2 + \text{M} \rightarrow \text{HO}_2 + \text{M}$	-47.10

Figure 1(a)

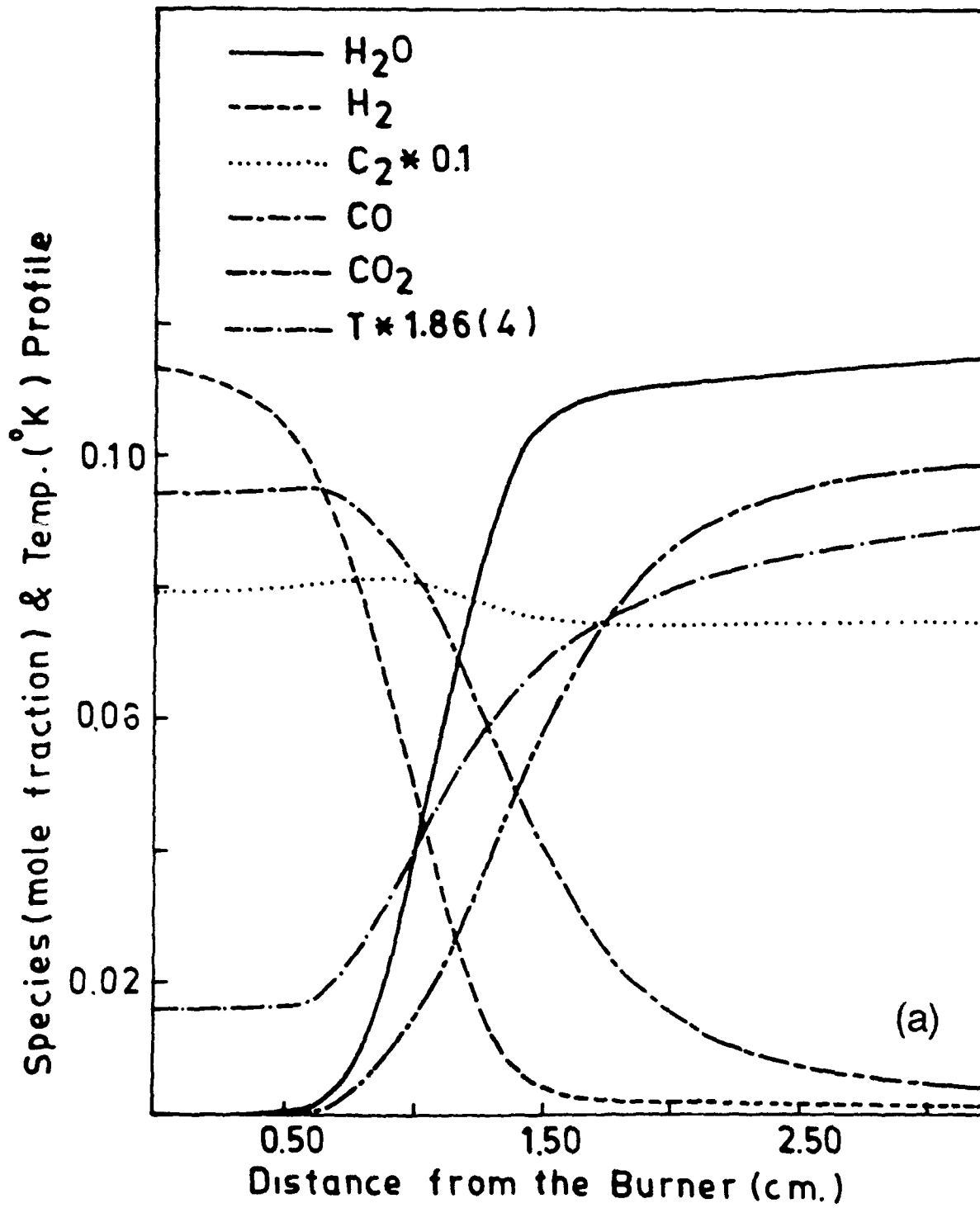


Figure 1(b)

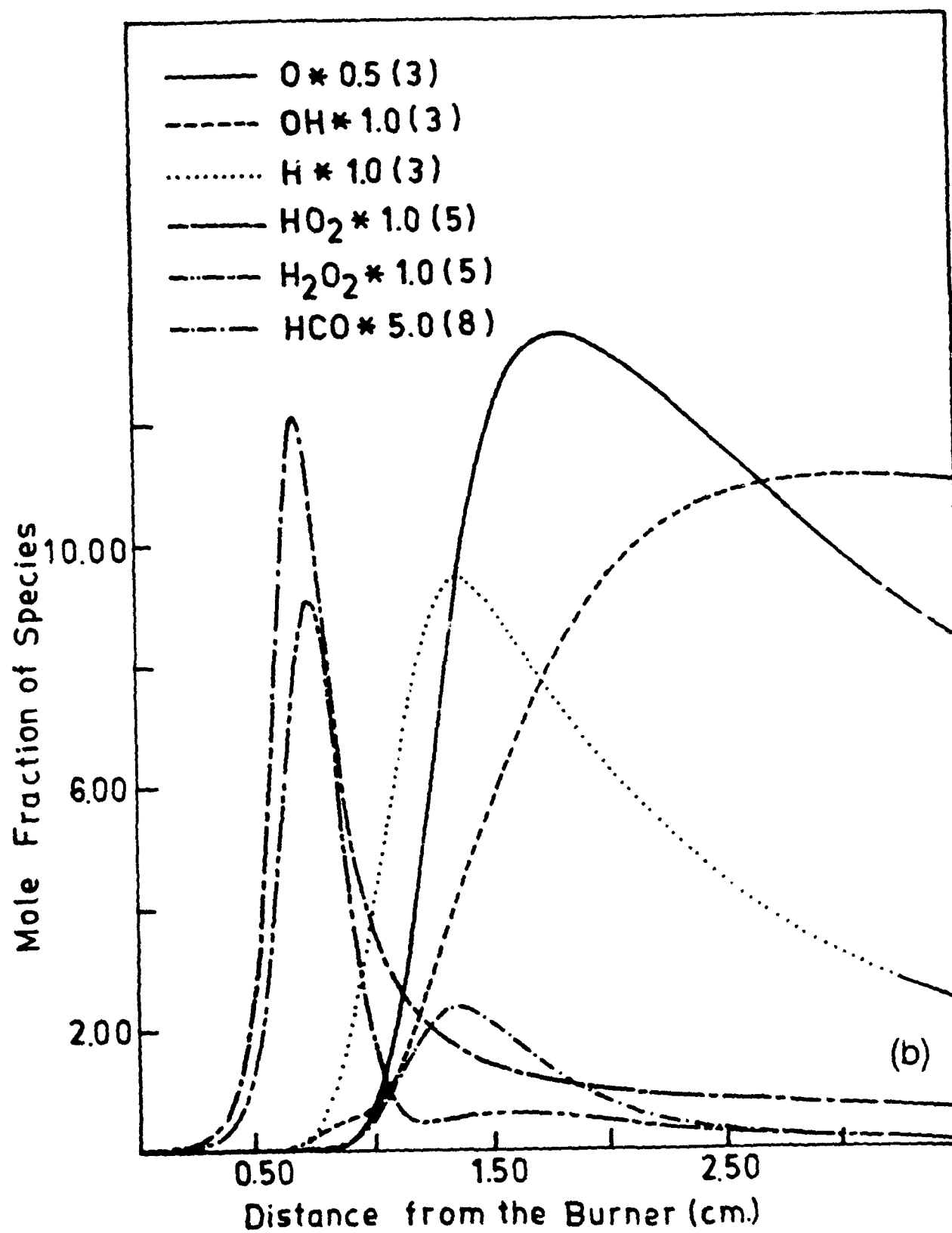


Figure 2(a)

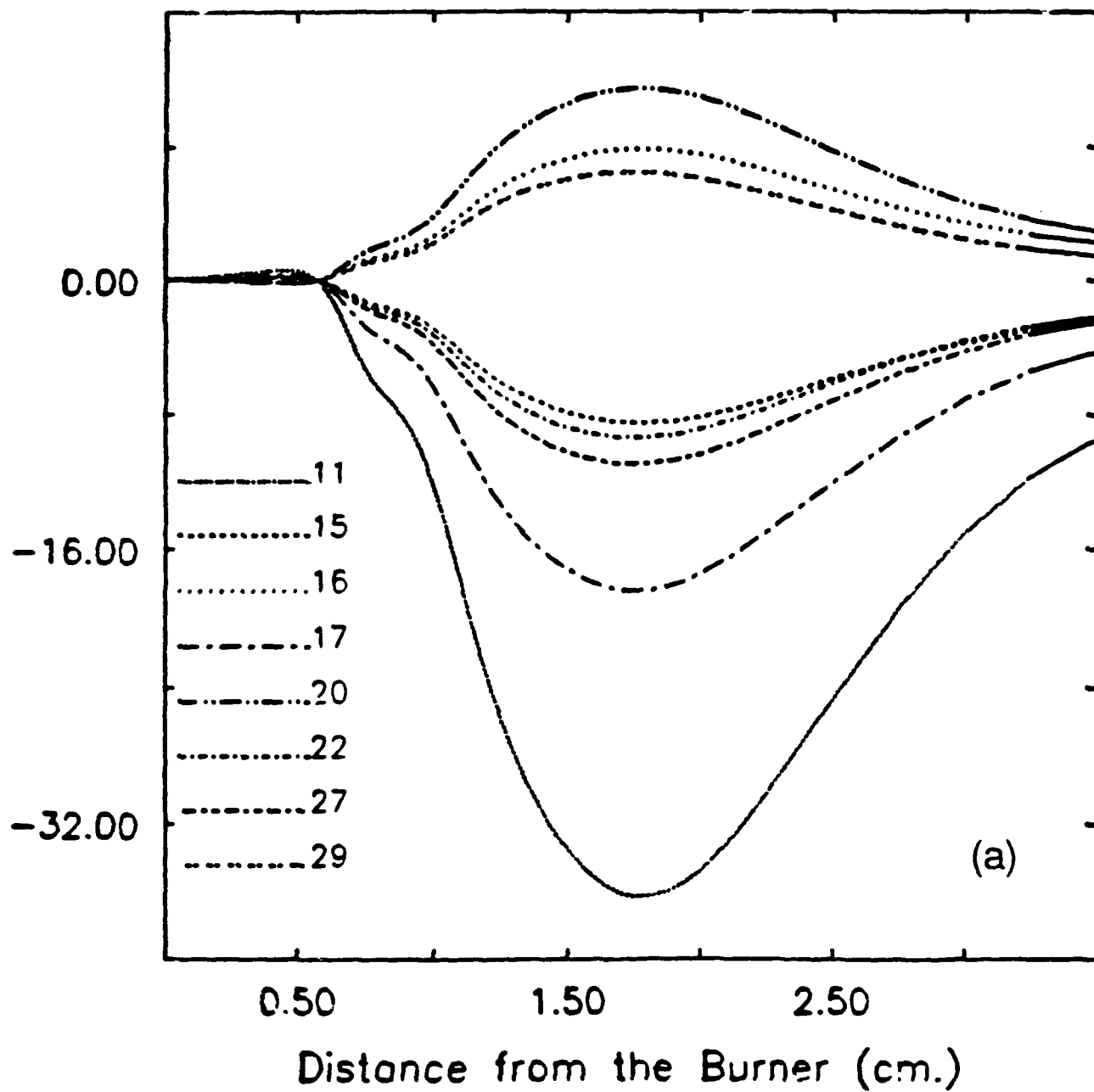


Figure 2(b)

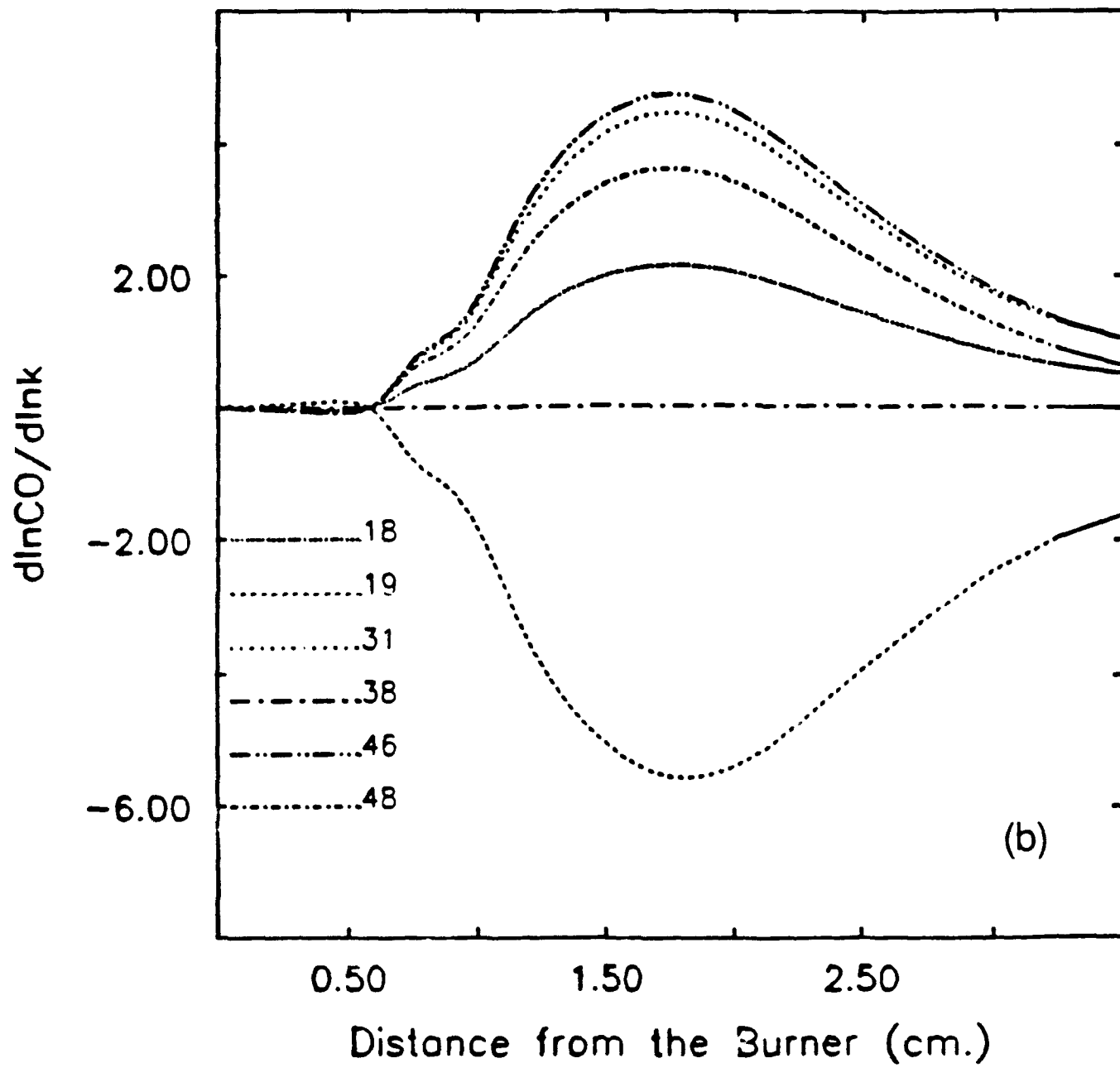


Figure 2(c)

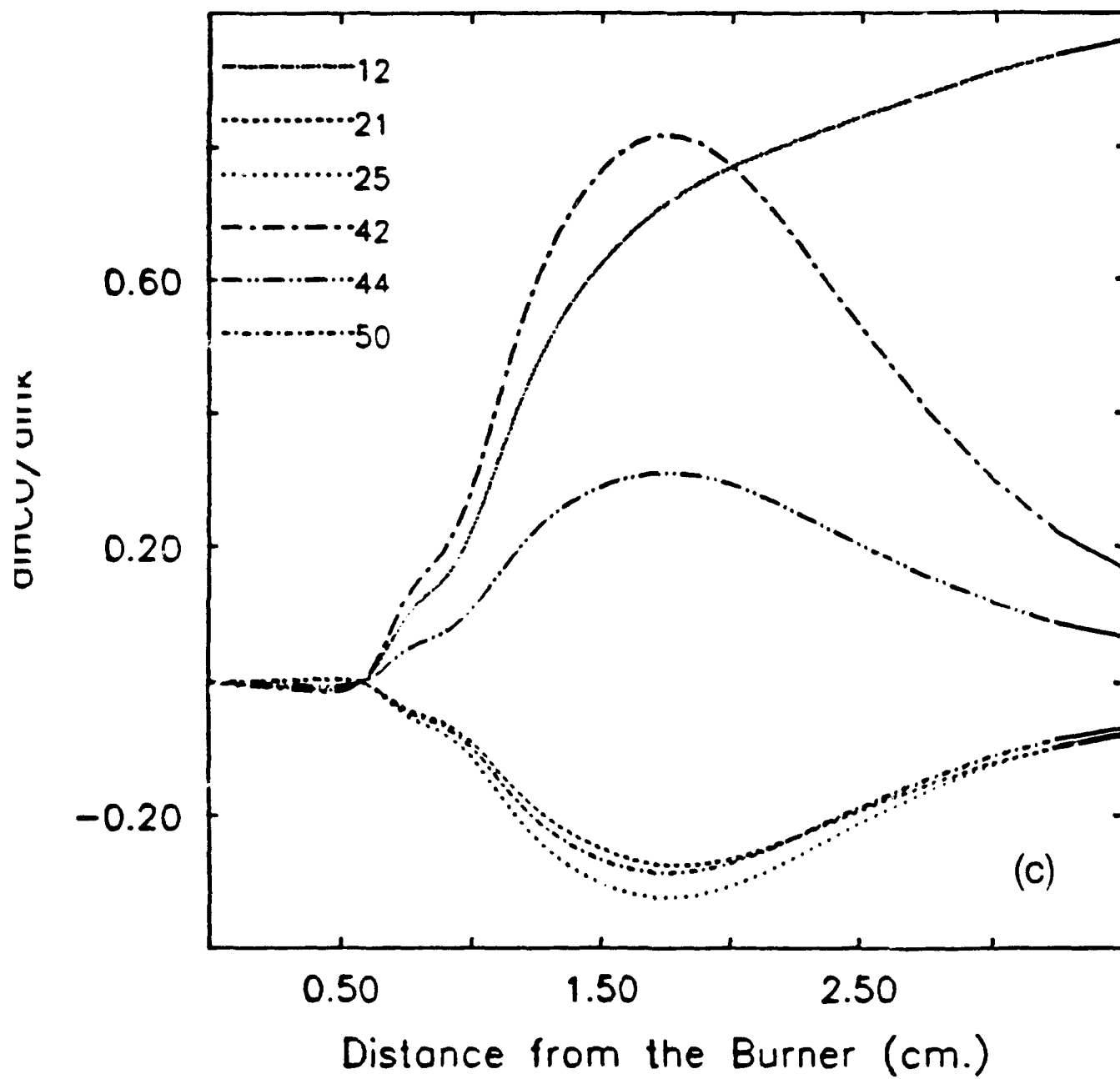


Figure 2(d)

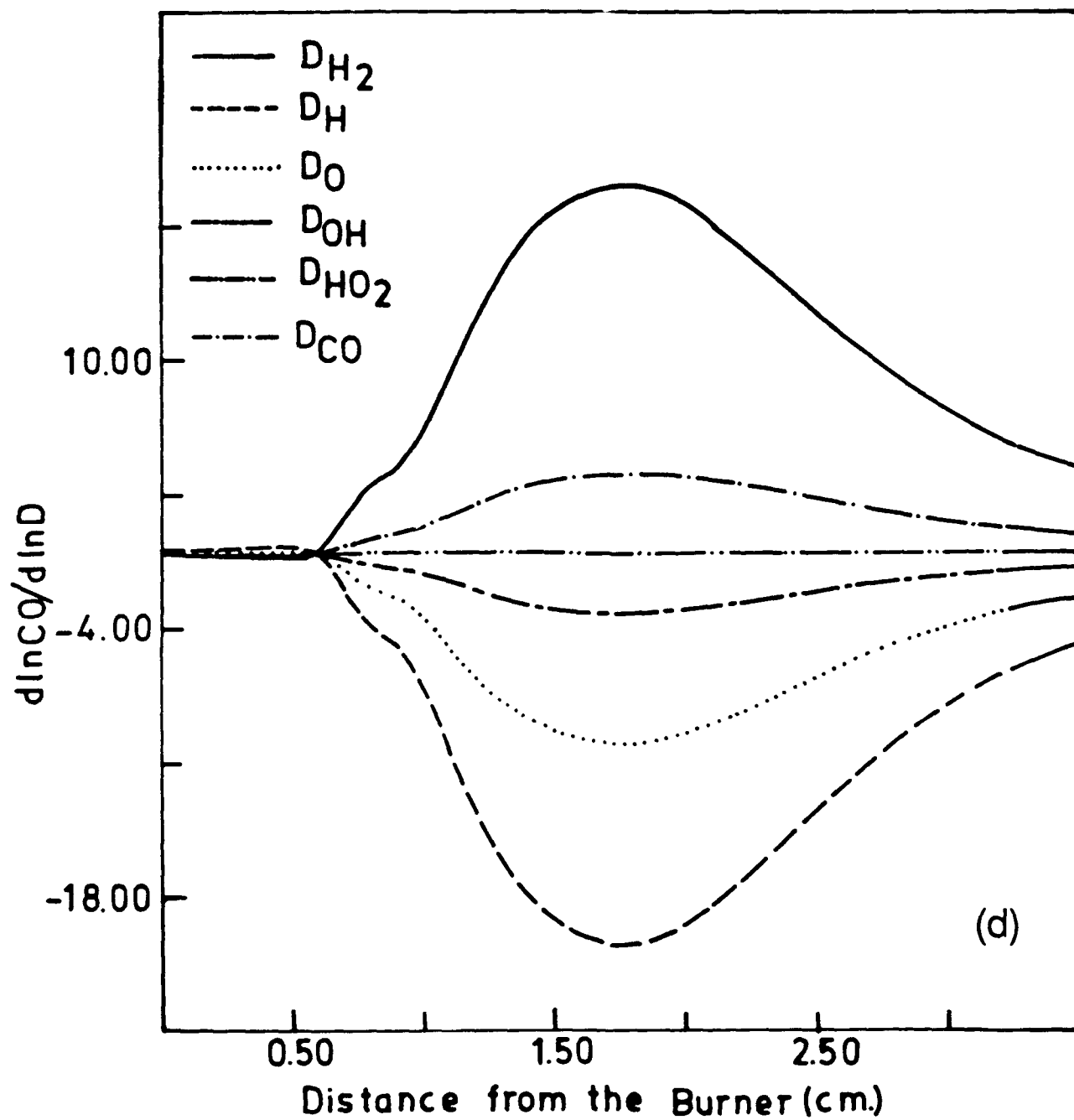


Figure 3

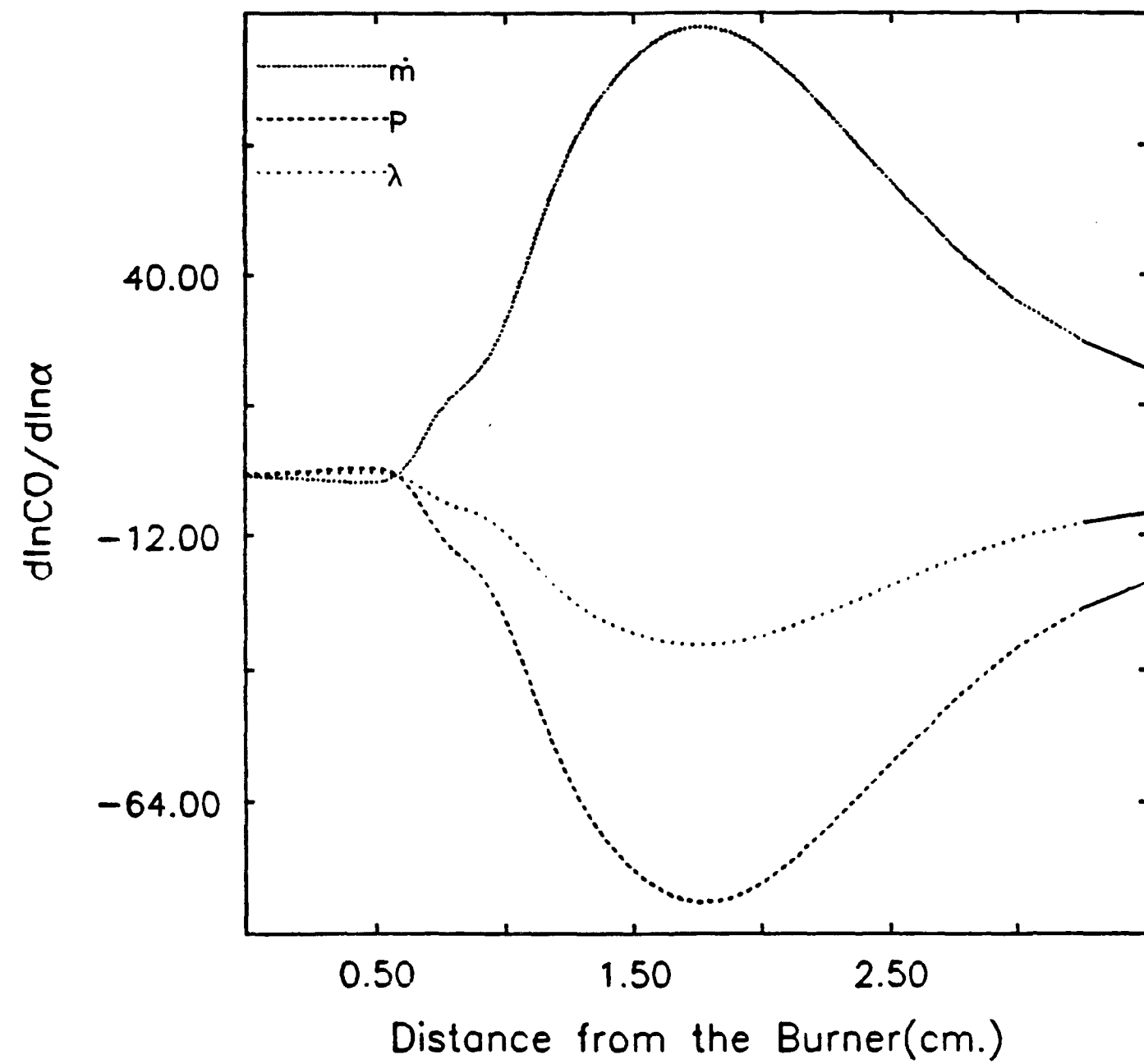


Figure 4

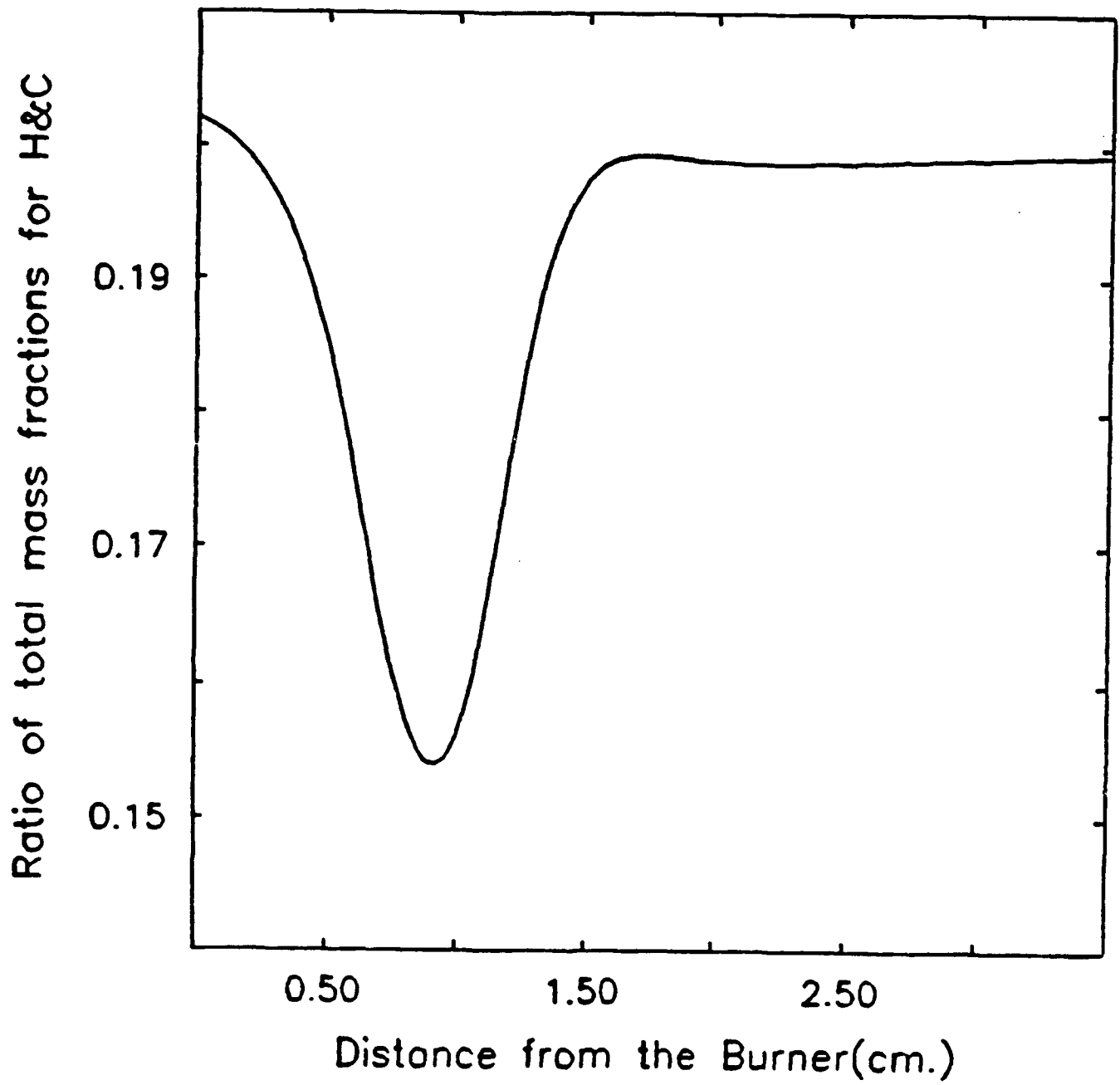


Figure 5(a)

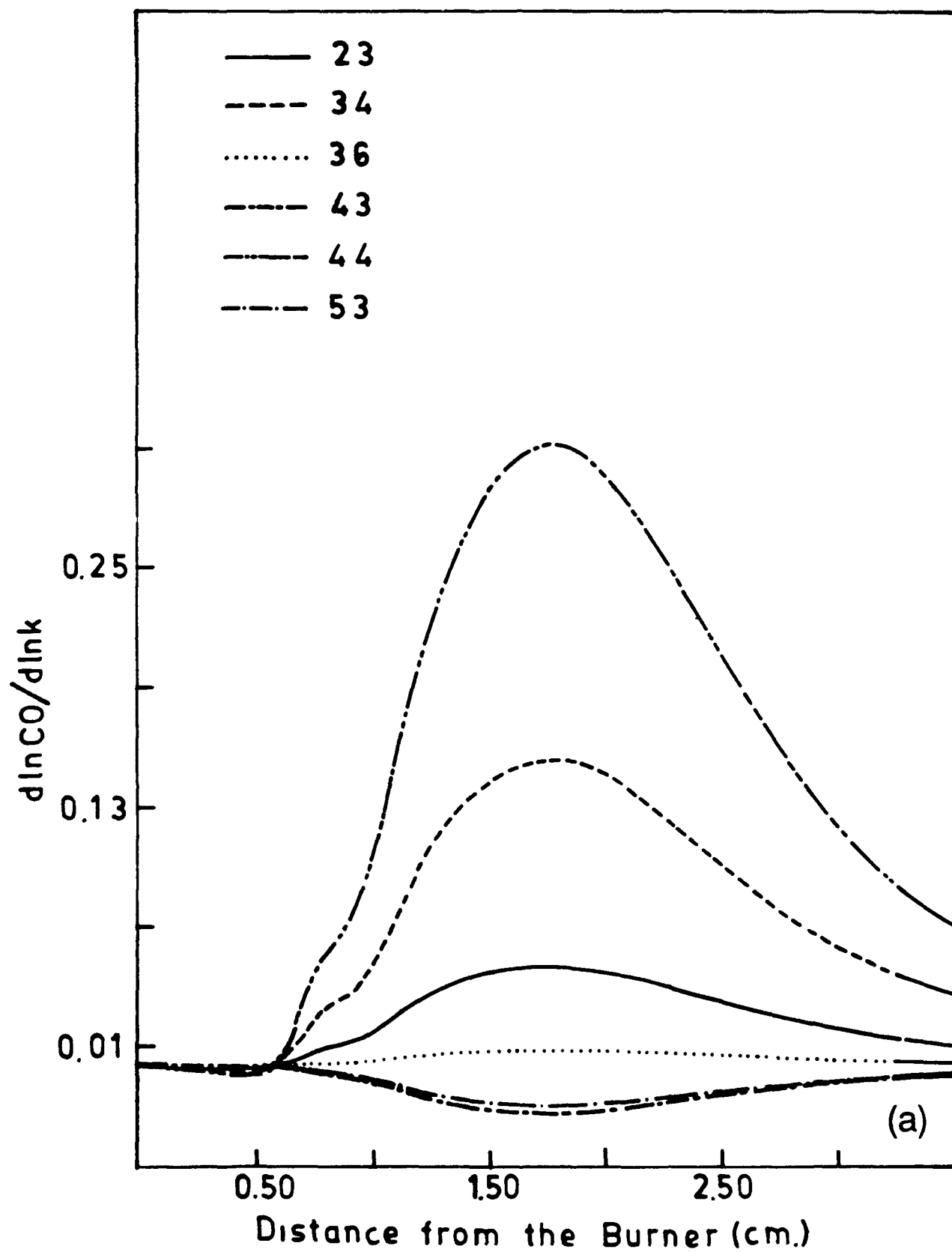


Figure 5(b)

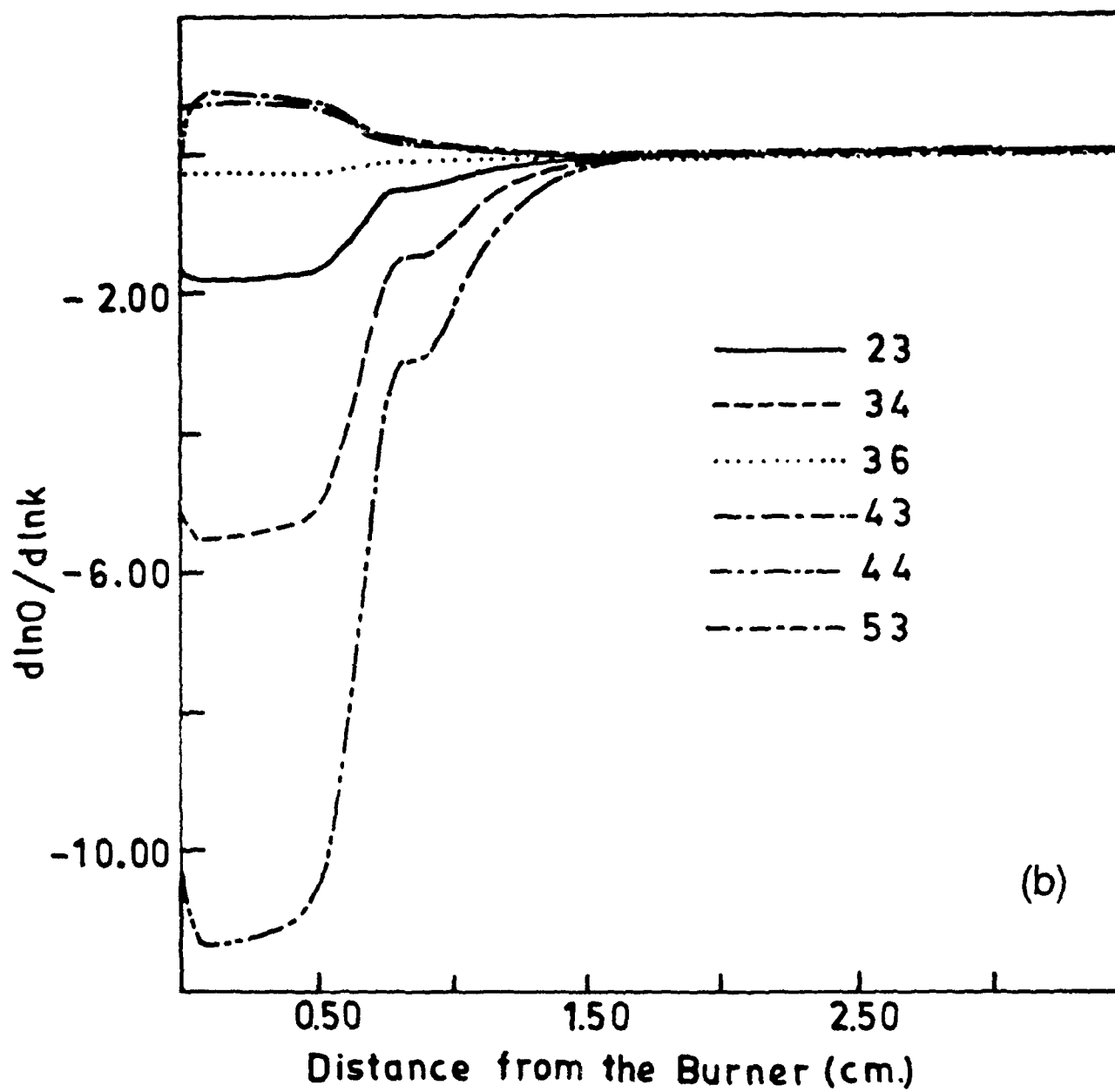


Figure 6

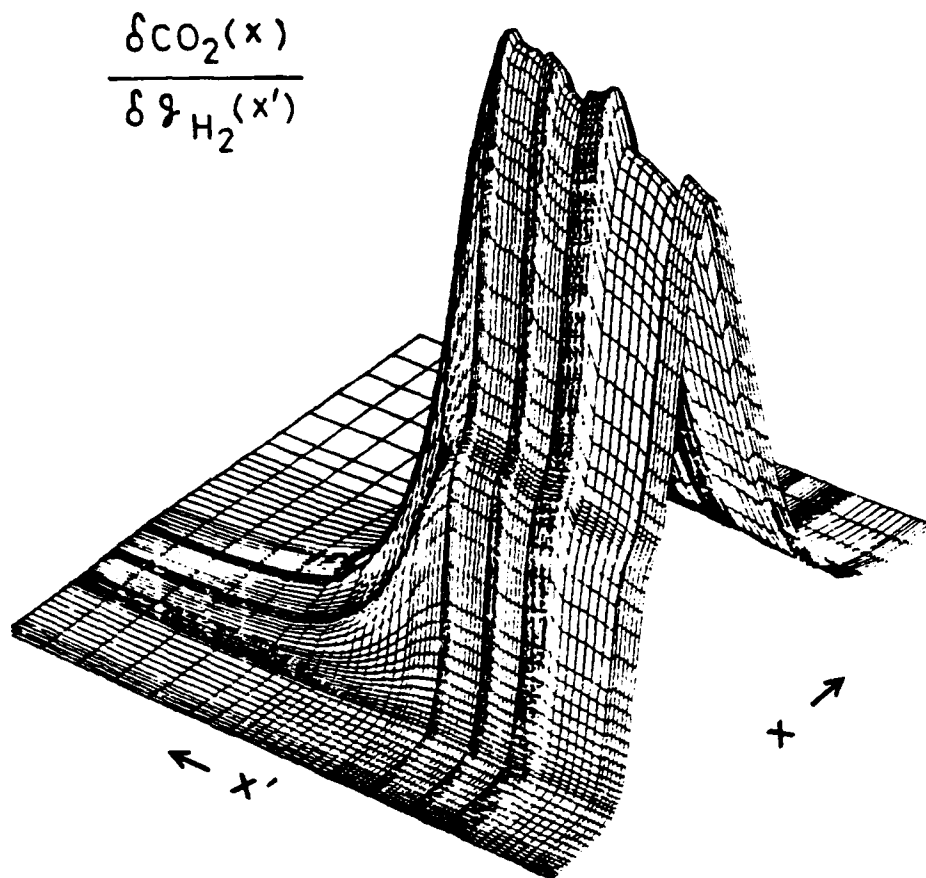


Figure 7(a)

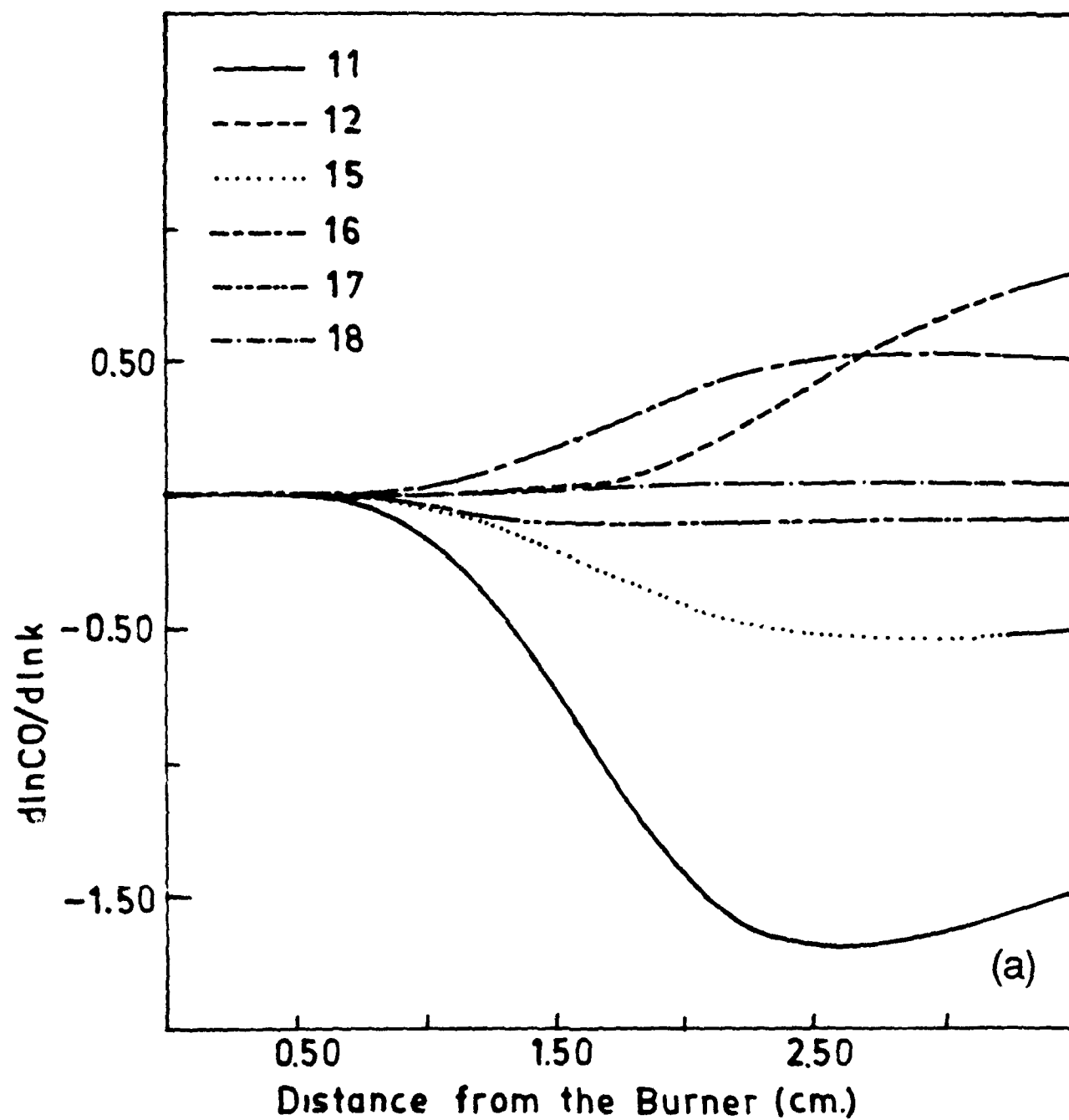


Figure 7(b)

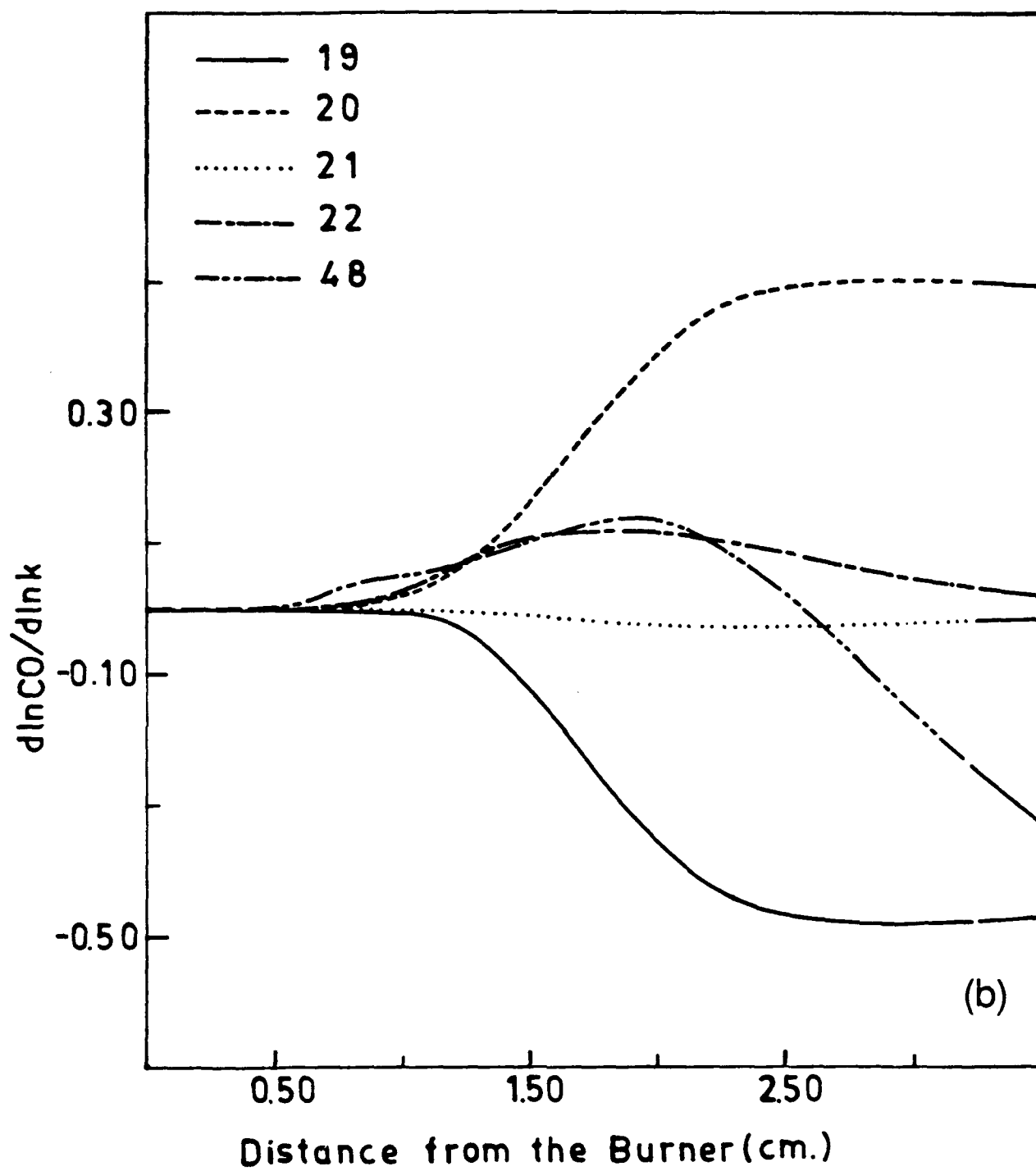
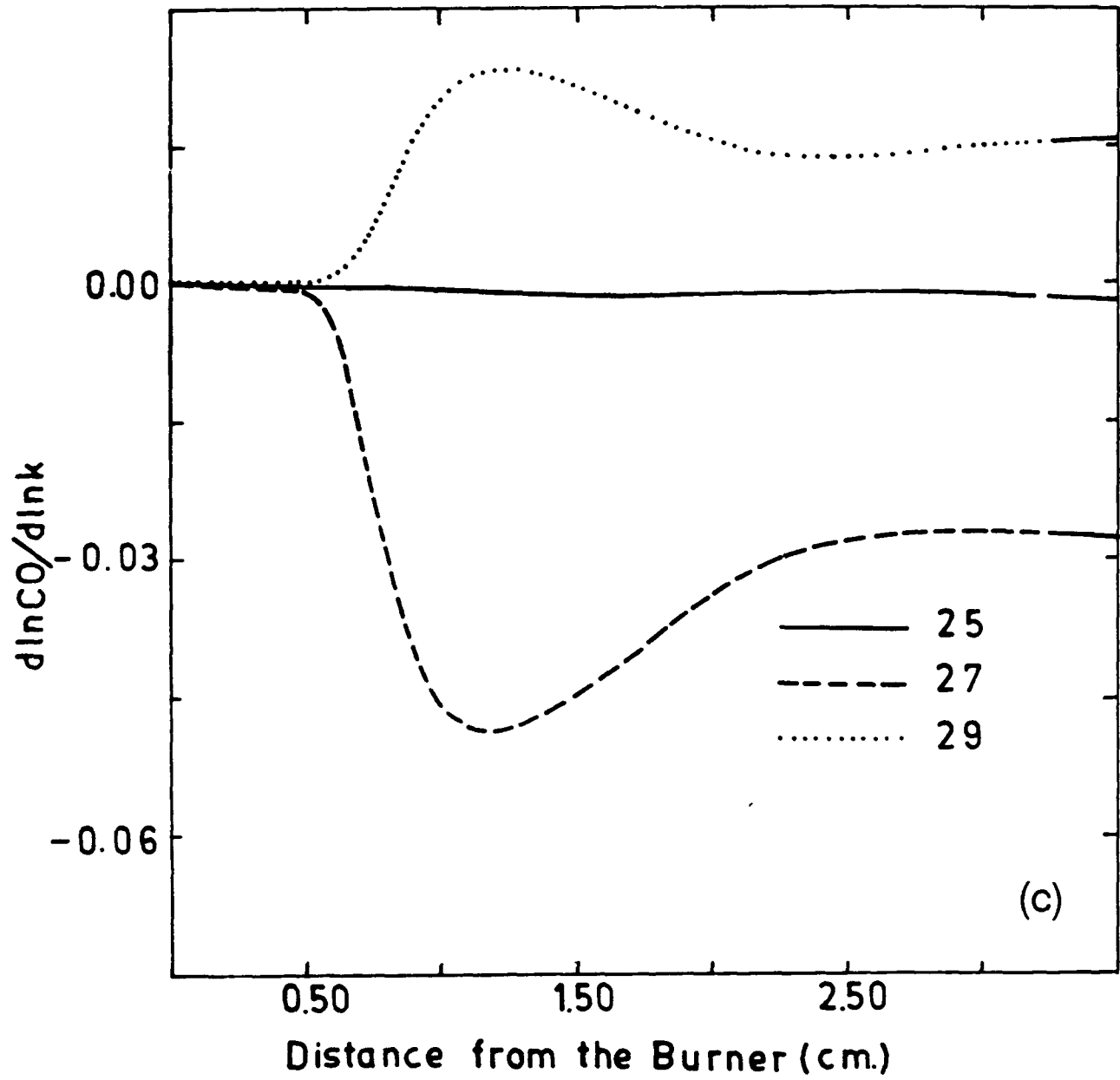


Figure 7(c)



Appendix F

6. A General Analysis of Approximate Lumping in Chemical Kinetics, G. Li and H. Rabitz, Chem. Eng. Sci., 45, 977 (1990).

A GENERAL ANALYSIS OF APPROXIMATE LUMPING IN CHEMICAL KINETICS

GENYUAN LI and HERSCHEL RABITZ*

Department of Chemistry, Princeton University, Princeton, NJ 08540, U.S.A.

(Received 19 December 1988; accepted 17 July 1989)

Abstract—A general analysis of approximate lumping is presented. This analysis can be applied to any reaction system with n species described by $dy/dt = f(y)$, where y is an n -dimensional vector in a desired region Ω , and $f(y)$ is an arbitrary n -dimensional function vector. Here we consider lumping by means of a rectangular constant matrix M (i.e. $\dot{y} = My$, where M is a row-full rank matrix and \dot{y} has dimension \tilde{n} not larger than n). The observer theory initiated by Luenberger is formally employed to obtain the kinetic equations and discuss the properties of the approximately lumped system. The approximately lumped kinetic equations have the same form $d\dot{y}/dt = Mf(\tilde{M}y)$ as that for exactly lumped ones, but depend on the choice of the generalized inverse \tilde{M} of M . The $\{1, 2, 3, 4\}$ -inverse is a good choice of the generalized inverse of M . The equations to determine the approximate lumping matrices M are presented. These equations can be solved by iteration. An approach for choosing suitable initial iteration values of the equations is illustrated by examples.

1. INTRODUCTION

A problem which frequently arises in the study of many subjects is the high dimensionality of mathematical models. Chemical problems of this type occur at the molecular level as well as in bulk kinetic phenomena. This paper will focus on kinetics where it is impractical and often not necessary to incorporate all the kinetic equations for each species in some complex reaction systems. Sometimes, even if the full set of kinetic equations are available, they are often needed in a reduced form for practical applications. Examples include day-to-day chemical plant operation or optimization for the design of an engine where integration of the full set of combustion partial differential equations would be prohibitive. Consequently, lumping, by which several species are combined as a single component, is often a necessity for theoretical and practical purposes. The theoretical analysis of lumping may also lead to some useful general conclusions. For example, the "principle of invariant response" obtained (Wei and Kuo, 1969) in the lumping analysis of unimolecular reaction systems has been used as a guidance for determination of the lumping scheme experimentally.

In a previous paper (Li and Rabitz, 1989) a general analysis of exact lumping was presented. Unfortunately, sometimes, even if a system is exactly lumpable, the resultant exact lumping schemes may not meet practically desired goals. For example, in the $\text{CO-H}_2\text{O-O}_2$ combustion system (Yetter *et al.*, 1985) we would like the easily measurable concentrations of CO , CO_2 , O_2 , and H_2O to be unlumped. With this constraint, the system likely cannot be exactly lumped, and we have to lump the species of the system

approximately. Developing a general approach for approximate lumping is very important for realistic practical problems. Approximate lumping has been discussed in some previous papers. Kuo and Wei (1969) proposed a method of constructing the lumped kinetic rate constant matrix for unimolecular reaction systems. Luss and Hutchinson (1971), Luss (1975), Golikeri and Luss (1972, 1974) and Hutchinson and Luss (1970) presented studies of the pitfalls and magnitude of errors in the use of empirical rate expressions for lumping many independent single or consecutive reactions. In the present paper we will treat the problem generally. Our exact lumping analysis will be employed as a rigorous starting point for the development of approximate lumping.

Section 2 of this paper presents the method to determine the kinetic equations of the approximately lumped system by the formal use of observer theory, and discussion is given on the properties of the lumped kinetic equations. The approximately lumped kinetic equations have the same form as those of exact lumping, but the error depends on the choices of the lumping matrix and its generalized inverse. In Section 3, the $\{1, 2, 3, 4\}$ -inverse will be proved to be a good choice of the generalized inverse of the lumping matrix and the equations to determine the approximate lumping matrix are derived. Section 4 considers the approximate lumping schemes valid in a given region of composition space. In Section 5, an approach for choosing suitable initial iteration values of the equations to determine the lumping matrix is presented. Section 6 presents some simple examples with the formulations. Finally, Section 7 gives a discussion of the results. The paper will draw heavily on the earlier work on exact lumping (Li and Rabitz, 1989), and the reader is guided to this reference for certain details.

* Author to whom correspondence should be addressed.

2. DETERMINING THE KINETIC EQUATIONS OF THE APPROXIMATELY LUMPED SYSTEM

2.4. Lumped system kinetics

Suppose the kinetics of an n -component reaction system can be described by

$$dy/dt = f(y) \quad (1)$$

where y is an n -composition vector and $f(y)$ is an arbitrary n -function vector which does not contain t explicitly.

Here we only consider a special class of lumping by means of an $\hat{n} \times n$ constant matrix M with rank \hat{n} ($\hat{n} \leq n$). If a system can be exactly lumped by the matrix M , it means that for

$$\hat{y} = My \quad (2)$$

we can find an \hat{n} -function vector $\hat{f}(\hat{y})$ such that

$$d\hat{y}/dt = \hat{f}(\hat{y}). \quad (3)$$

If a system is not exactly lumpable by a given M , one cannot find a set of differential equations as eq. (3) to describe the behaviour of \hat{y} . In this case we need to find a set of differential equations to describe the behavior of \hat{y} approximately. Liu and Lapidus (1973) formally employed the observer theory initiated by Luenberger (1964) for control problems to obtain the necessary and sufficient conditions of exact and approximate lumping for unimolecular reaction system. Here we further extend this approach to nonlinear systems for the determination of the kinetic equations of the approximately lumped system. Although no actual observations are assumed to have been made, the analogy with observer theory is nevertheless still useful.

The output $y(t)$ of the kinetic system in eq. (1) can be employed to drive another system described by

$$d\hat{y}/dt = \hat{f}(\hat{y}) + e(y) \quad (4)$$

where $e(y)$ is an \hat{n} -dimensional function vector called the error vector. The second system in eq. (4) is the observer of the first one in eq. (1). Then we have the following statement: let S_1 be an n -component kinetic system described by eq. (1), which drives another \hat{n} th-order ($\hat{n} \leq n$) lumped kinetic system S_2 described by eq. (4). Suppose there is an $\hat{n} \times n$ row-full rank constant lumping matrix M satisfying

$$Mf(y) = \hat{f}(\hat{y}) + e(y). \quad (5)$$

If $\hat{y}(0) = My(0)$, then $\hat{y}(t) = My(t)$ for all $t \geq 0$, or more generally:

$$My(t) - \hat{y}(t) = \text{constant}. \quad (6)$$

This statement can be proved as follows. Suppose that such a lumping matrix did exist, i.e. it satisfies eq. (6). The two systems are governed by eqs (1) and (4). Using eq. (6) we have

$$d[My(t)]/dt = d\hat{y}(t)/dt.$$

Considering eqs (1) and (4) one obtains

$$Mf(y) = \hat{f}(\hat{y}) + e(y).$$

This condition is also sufficient; if there exists a matrix M satisfying eq. (5), it will be shown that M has the property of this statement. Using eqs (1) and (4) we have

$$Mdy/dt - d\hat{y}/dt = Mf(y) - \hat{f}(\hat{y}) - e(y)$$

$$d(My - \hat{y})/dt = 0$$

i.e.

$$My(t) - \hat{y}(t) = \text{constant}. \quad (7)$$

When $\hat{y}(0) = My(0)$, the constant vector is the null vector. Then we have

$$\hat{y}(t) = My(t).$$

For given M and $f(y)$ it is always possible to construct a pair of $\hat{f}(\hat{y})$ and $e(y)$ to satisfy eq. (5). Therefore, we can always find a set of differential equations as eq. (4) to describe the behavior of the lumped species \hat{y} . We can see that exact lumping is just the special case $e(y) = 0$. In the exact case eq. (5) becomes

$$Mf(y) = \hat{f}(\hat{y}) \quad (8)$$

which was given in our previous paper about analysis of exact lumping.

From eq. (5) we see for a given M and $\hat{f}(\hat{y})$ that $e(y)$ is uniquely determined by

$$e(y) = Mf(y) - \hat{f}(My). \quad (9)$$

However, for a given M and $e(y)$, $\hat{f}(\hat{y})$ may not exist. For example if $e(y)$ is taken to be the identically zero function, the appropriate $\hat{f}(\hat{y})$ exists only if the original system is exactly lumpable by M . A reasonable expectation is that $\hat{f}(\hat{y})$ have the same form to that of the exactly lumped equations:

$$\hat{f}(\hat{y}) = Mf(\bar{M}\hat{y}) \quad (10)$$

where \bar{M} is a $\{1, 2, 3\}$ -generalized inverse of M (Ben-Israel and Greville, 1974) satisfying

$$M\bar{M} = I_{\hat{n}}. \quad (11)$$

Under this condition we can prove that $e(y)$ satisfies

$$e(\bar{M}My) = 0. \quad (12)$$

Indeed, if we choose $\hat{y}(0) = My(0)$, then we obtain $\hat{y}(t) = My(t)$. Substituting eq. (10) into eq. (5) and rearranging it yields

$$\begin{aligned} e(y) &= Mf(y) - Mf(\bar{M}\hat{y}) \\ &= Mf(y) - Mf(\bar{M}My). \end{aligned} \quad (13)$$

This is valid for any value of y . Therefore, since y can be arbitrary we choose $y = \bar{M}My$, then

$$\begin{aligned} e(\bar{M}My) &= Mf(\bar{M}My) - Mf(\bar{M}M\bar{M}My) \\ &= Mf(\bar{M}My) - Mf(\bar{M}My) \\ &= 0. \end{aligned} \quad (14)$$

For exact lumping $\hat{f}(\hat{y})$ is unique and does not depend on the choice of \bar{M} . However, now this is no longer true. Both $e(y)$ and $\hat{f}(\hat{y})$ are dependent on the choice of \bar{M} . Under the constraint of eq. (10), eq. (4)

can be represented as

$$d\hat{y}/dt = Mf(\bar{M}\hat{y}) + e(y). \quad (15)$$

Equation (15) does not actually reduce the dimension of the system, because the 1st term $e(y)$ is a function of y . However, if the term $e(y)$ for given M and \bar{M} is small compared to the first term on the right-hand side of eq. (15), and does not significantly effect the solution, the lumped system can be approximately described by

$$d\hat{y}/dt = Mf(\bar{M}\hat{y}). \quad (16)$$

In order to minimize $e(y)$ our task is to develop an approach to determine appropriate M and \bar{M} . Notice that \hat{y} in eq. (16) is equal to My only if the original system is exactly lumpable by M . For the sake of simplicity, we do not distinguish the \hat{y} in eq. (16) and the $\hat{y} = My$, but the reader should keep this in mind.

2B. The properties of $\hat{f}(\hat{y})$ and $e(y)$

The conditions $e(\bar{M}My) = 0$ has some special properties. The mapping by the projection operator $\bar{M}M$ becomes an "endomorphism" of the composition Y_n -space. The range of this endomorphic mapping is an \hat{n} -dimensional $Y_{\hat{n}}$ -subspace of the composition space. The equation $e(\bar{M}My) = 0$ means that for any value of y in the $Y_{\hat{n}}$ -subspace $\hat{f}(\hat{y})$ is exactly equal to $Mf(y)$ and the system is then exactly lumpable in this region.

Suppose the original kinetic system has a stable point y^* for a given initial composition such that

$$\lim_{t \rightarrow \infty} y(t) = y^*. \quad (17)$$

This is a common circumstance for most kinetic systems. If we can choose the generalized inverse \bar{M} of M such that the stable point y^* is in the $Y_{\hat{n}}$ -subspace, then we have

$$e(\bar{M}My^*) = e(y^*) = 0. \quad (18)$$

Let $\hat{y}^* = My^*$ and substitute it into eq. (15). Considering that $f(y^*) = 0$, we obtain

$$\begin{aligned} d\hat{y}^*/dt &= Mf(\bar{M}\hat{y}^*) + e(y^*) \\ &= Mf(\bar{M}My^*) + e(\bar{M}My^*) \\ &= Mf(y^*) = 0. \end{aligned}$$

This indicates that in this case $\hat{y}^* = f(y^*)$ is the stable point of eq. (16). When both the original and the lumped systems have only one stable point, the above discussion implies that, when t becomes larger and larger, the solution of eq. (16) will be closer and closer to the exact solution of eq. (1). A similar observation for unimolecular reaction systems was obtained by Kuo and Wei (1959).

The determination of $\hat{f}(\hat{y})$ by eq. (10) has another property. Since

$$\hat{f}(\hat{y}) = Mf(\bar{M}\hat{y})$$

$\hat{f}_i(\hat{y})$ is a linear combination of the elements of $f(\bar{M}\hat{y})$. Therefore $\hat{f}(\hat{y})$ can be determined not only directly

from $f(y)$ but usually has a form similar to that of $f(y)$ as well.

3. THE EQUATIONS FOR DETERMINING THE APPROXIMATE LUMPING SCHEMES

From eq. (13) one can see that $e(y)$ is a function of M and \bar{M} . Therefore, if we desire to use eq. (16) as an approximately lumped model, we need to determine suitable M and \bar{M} , which give the smallest $e(y)$ in the desired region of Y_n -space. There may be several ways to reach this goal; however, since we use the same formula for approximate lumping as that of the exact case, we will apply our results of exact lumping as a starting point to solve this problem.

3A. Exact and approximate lumping in a desired region of the composition Y_n -space

In realistic problems the lumping schemes are usually desired in a particular region Ω of the composition Y_n -space. In the previous paper on exact lumping, we did not give any restriction on the values of y , i.e. y can take any value in Y_n . When y is required in a desired region Ω , we will demonstrate that the necessary and sufficient condition for the existence of exact lumping of eq. (1) are the same except that $y \in \Omega$: (1) the subspace \mathcal{H} spanned by the row vectors of the lumping matrix M is a fixed invariant one under $J^T(y)$ for all values of $y \in \Omega$, and (2) M satisfies the following equation

$$M[J(y) - J(\bar{M}My)] = 0 \quad \forall y \in \Omega. \quad (19)$$

Let Ω_i represent the region of $\bar{M}_i My$, where $y \in \Omega$ and \bar{M}_i is a particular generalized inverse of M satisfying $M\bar{M}_i = I_{\hat{n}}$. First, we will prove that these two conditions hold for all $y \in \Omega_i$ if they hold in Ω . Since \mathcal{H} is $J^T(y)$ -invariant in Ω , eq. (19) can be rewritten as

$$Q(y)M = M J^T(\bar{M}_i My) \quad \forall y \in \Omega \quad (20)$$

where $Q(y)$ is an unspecified $\hat{n} \times \hat{n}$ matrix. This implies that \mathcal{H} is also $J^T(y)$ -invariant in the region Ω_i . Letting $\tilde{y} = \bar{M}_i My$, then we obtain

$$\begin{aligned} M[J(\tilde{y}) - J(\bar{M}_i M\tilde{y})] &= M[J(\bar{M}_i My) \\ &\quad - J(\bar{M}_i M\bar{M}_i My)] = M[J(\bar{M}_i My) \\ &\quad - J(M_i My)] = 0. \end{aligned} \quad (21)$$

Thus eq. (19) is also valid in Ω_i .

For a given M there are an infinite number of \bar{M} . The general form of them is

$$\bar{M} = \bar{M}_i + (I_n - \bar{M}_i M)Z \quad (22)$$

where \bar{M}_i is any given generalized inverse of M satisfying $M\bar{M}_i = I_{\hat{n}}$, and Z is an arbitrary $n \times \hat{n}$ matrix. The reader can readily prove that \bar{M} given by eq. (22) satisfies $M\bar{M} = I_{\hat{n}}$ and any \bar{M}_j satisfying $M\bar{M}_j = I_{\hat{n}}$ can be represented in the form of eq. (22) as follows:

$$\bar{M}_j = \bar{M}_i + (I_n - \bar{M}_i M)(\bar{M}_j - \bar{M}_i). \quad (23)$$

Taking account of eqs (12) and (15) we know that in Ω_i the system described by eq. (1) is always exactly

lumpable by M . Since the two conditions hold for any Ω_i [its \bar{M}_i satisfies eq. (19)] if they hold in Ω , we can therefore consider exact lumping in the region $\Omega_{\text{total}} = \cup_{i=1}^r \Omega_i$ (where $\Omega_c = \Omega$) instead of Ω . Following the same procedure as in our previous paper on exact lumping one can prove that the two conditions are necessary. We will demonstrate that they are also sufficient.

Notice that if Ω is connected, so is $\cup_{i=1}^r \Omega_i$. This is because that the elements of \bar{M} s can change continuously by continuously changing the elements of Z . Then the images of $\bar{M}M\Omega$ are continuous and connected. We can further prove that Ω is also connected with $\cup_{i=1}^r \Omega_i$. Suppose that there is a vector $y \notin \text{Ker } M$ in Ω (otherwise My in Ω are identically zero and there is no necessity to consider lumping). Then we can demonstrate that there exists a projection operator $P = \bar{M}M$ in Y_n with $Py = y$.

Since $My = c \neq 0$, one can always find a nonsingular $n \times n$ matrix Q such that

$$QMy = M'y = Qc = e_1 \quad (24)$$

where M' is another matrix representation of M and e_1 is the unit vector. There also exist $n-1$ vectors w_i satisfying

$$M'w_i = e_{i+1} \quad (i = 1, 2, \dots, n-1). \quad (25)$$

Let y and w_i s compose the matrix

$$\bar{M}' = (y \ w_1 \ \dots \ w_{n-1}). \quad (26)$$

Then we have

$$M'\bar{M}' = I_n. \quad (27)$$

The matrix $P = \bar{M}'M'$ is a projection operator due to $P^2 = P$ and

$$Py = \bar{M}'M'y = \bar{M}'e_1 = y. \quad (28)$$

Letting

$$\bar{M} = \bar{M}'Q \quad (29)$$

yields

$$M\bar{M} = Q^{-1}M'\bar{M}'Q = I_n \quad (30)$$

and

$$P = \bar{M}'M' = \bar{M}Q^{-1}QM = M\bar{M}. \quad (31)$$

This result shows that we can find a generalized inverse \bar{M} of M such that $\bar{M}My = y$. This implies that $\Omega \cap \cup_{i=1}^r \Omega_i \neq \emptyset$, and then the whole Ω_{total} is connected.

From the above two necessary conditions of exact lumping in Ω we can deduce the following equation:

$$MJ(y) = MJ(\bar{M}My)\bar{M}M \quad \forall y \in \Omega_{\text{total}}. \quad (32)$$

Since Ω_{total} is connected, we can choose a trajectory starting from a point y_0 in $\bar{M}M\Omega$ (where \bar{M} is any of the generalized inverses \bar{M}_i) to an arbitrary point y in Ω and integrate eq. (32) with respect to y along this

trajectory:

$$\begin{aligned} \int_{y_0}^y M[J(y) - J(\bar{M}My)\bar{M}M] dy \\ = M[f(y) - f(\bar{M}My)] - M[f(y_0) - f(\bar{M}My_0)] \\ = M[f(y) - f(\bar{M}My)] = 0. \end{aligned} \quad (33)$$

Here we used the relation $y_0 = \bar{M}My_0$. Equation (33) gives

$$Mf(y) = Mf(\bar{M}My) \quad \forall y \in \Omega. \quad (34)$$

Then the system described by eq. (1) with the constraint $y \in \Omega$ can be exactly lumped by M and the lumped kinetic equations are eq. (16). Therefore, these conditions are sufficient.

The first necessary and sufficient condition of exact lumping in Ω can be represented as the following equation

$$(I_n - M^T\bar{M}^T)J^T(y)M^T = 0 \quad \forall y \in \Omega. \quad (35)$$

This follows because the null space of the projection operator $I_n - M^T\bar{M}^T$ is \mathcal{H} . If \mathcal{H} is a fixed $J^T(y)$ -invariant subspace (independent on $y \in \Omega$), i.e.

$$J^T(y)M^T = M^TQ^T(y) \quad (36)$$

then

$$\begin{aligned} (I_n - M^T\bar{M}^T)J^T(y)M^T &= (I_n - M^T\bar{M}^T)M^TQ^T(y) \\ &= (M^T - M^T)Q^T(y) \\ &= 0. \end{aligned} \quad (37)$$

If a system is not exactly lumpable, eqs (19) and (35) do not hold exactly. In this case, it is natural to define two error matrices $E_1(y)$ and $E_2(y)$ to describe the deviation from the exact lumping for given M and \bar{M} :

$$E_1(y) = (I_n - M^T\bar{M}^T)J^T(y)M^T \quad \forall y \in \Omega \quad (38)$$

$$E_2(y) = M[J(y) - J(\bar{M}My)] \quad \forall y \in \Omega. \quad (39)$$

For approximate lumping, our task is simply to find appropriate M and \bar{M} , which will minimize the absolute values of all elements of $E_1(y)$ and $E_2(y)$ in the desired region of y . We will first determine \bar{M} in Section 3B based on minimization of $E_1(y)$, and then present the equations to determine M by minimization of $E_1(y)$ or $E_2(y)$ for all values of y in Ω in Sections 3C and D, respectively. Finally, at the end of Section 3D the simultaneous minimization of $E_1(y)$ and $E_2(y)$ to get M will be discussed.

3B. Determination of the generalized inverse \bar{M}

For a given M there are an infinite number of \bar{M} , which makes the determination of approximate lumping schemes very complicated. Several considerations on the choice of \bar{M} might be made for different purposes. For example, possible requirements are that the lumped model follows a uni- and or bimolecular reaction scheme and that the image of the equilibrium point of the original system upon mapping by $\bar{M}M$ is in the Y_n -subspace. In this case, \bar{M} must satisfy other restrictions. Here we only consider the determination

of \bar{M} by a minimum demand, i.e. the smallest $E_1(y)$ and $E_2(y)$ for a given M . We will prove that the $\{1, 2, 3, 4\}$ -inverse will give the smallest $E_1(y)$ for any value of y . When \mathcal{H} consists of an orthonormal basis, i.e.

$$MM^T = I_n \quad (40)$$

the $\{1, 2, 3, 4\}$ -inverse is simply M^T . Then the determination of M and \bar{M} will be reduced to only determining M . In order to represent E_1 as being the function of \bar{M} and y for a given M , we will use the symbol $E_1(\bar{M}, y)$ here.

It is reasonable to denote a measure of the error $Z(\bar{M}, y)$ for a given \bar{M} and y by the trace of matrix $E_1^T(\bar{M}, y)E_1(\bar{M}, y)$, which is the sum of the squares of all the elements in $E_1(\bar{M}, y)$:

$$Z(\bar{M}, y) = \text{tr } E_1^T(\bar{M}, y)E_1(\bar{M}, y). \quad (41)$$

Our task is to choose an \bar{M} such that $Z(\bar{M}, y)$ has the smallest possible value in a desired region Ω of y .

As a mathematical preliminary observe that an $n \times n$ symmetric nonnegative definite matrix B , denoted as $B \geq 0$, can be represented as PP^T and $\text{tr } B \geq 0$. If both A and B are $n \times n$ symmetric nonnegative definite matrices with $A - B \geq 0$, then we say that $A \geq B \geq 0$. Thus from

$$\text{tr } A - \text{tr } B = \text{tr } (A - B) \geq 0$$

we have

$$\text{tr } A \geq \text{tr } B \geq 0. \quad (42)$$

We can now make use of this relation to find the best choice of \bar{M} . Letting M^* represent the $\{1, 2, 3, 4\}$ -inverse of M and considering eq. (38) followed by algebraic manipulations one may establish that

$$\begin{aligned} E_1^T(\bar{M}, y)E_1(\bar{M}, y) - E_1^T(M^*, y)E_1(M^*, y) \\ = MJ(y)(I_n - \bar{M}M)(I_n - M^T\bar{M}^T)J^T(y)M^T \\ - MJ(y)(I_n - M^*M)(I_n - M^TM^*{}^T)J^T(y)M^T \\ = [MJ(y)(M^* - \bar{M})M][MJ(y) \\ \times (M^* - \bar{M})M]^T \geq 0. \end{aligned} \quad (43)$$

Here we used the properties of the $\{1, 2, 3, 4\}$ -inverse (Ben-Israel and Greville, 1974), i.e.

$$\begin{aligned} MM^*M &= M & M^*MM^* &= M^* \\ (MM^*)^T &= MM^* & (M^*M)^T &= M^*M. \end{aligned} \quad (44)$$

For brevity we leave the proof of eq. (43) to the reader. Since $E_1^T(\bar{M}, y)E_1(\bar{M}, y)$ and $E_1^T(M^*, y)E_1(M^*, y)$ are nonnegative definite, we may use eq. (42) to show that

$$Z(\bar{M}, y) \geq Z(M^*, y) \geq 0 \quad \forall \bar{M} \in M\bar{M} = I_n.$$

Notice that there is no restriction on y so it is valid for any value of $y \in \Omega$. Therefore $\bar{M} = M^*$ gives the global minimum of E_1 for a given M . Since the error of lumping is independent of the choice of the basis for a given fixed $J^T(y)$ -invariant subspace \mathcal{H} , then we let M satisfy eq. (40) and adopt the choice $\bar{M} = M^T$. We should emphasize that this choice may not be perfect, because E_2 is not considered.

3C. The matrix equations for determining M under minimization of E_1

After choosing $\bar{M} = M^T$, we only need to determine M , which will minimize $E_1(y)$ and $E_2(y)$ in the desired region of y . In this case $E_1(y)$ is represented as

$$E_1(y) = (I_n - M^TM)J^T(y)M^T. \quad (45)$$

Here we discuss two cases: unconstrained and constrained approximate lumping matrices.

(1) Unconstrained approximate lumping matrices.

Just like the determination of \bar{M} , we define the error $Z_1(y)$ for given M and y by the trace of the matrix $E_1^T(y)E_1(y)$, which is the sum of the squares of all the elements in $E_1(y)$:

$$\begin{aligned} Z_1(y) &= \text{tr } [E_1^T(y)E_1(y)] \\ &= \text{tr } [MJ(y)(I_n - M^TM)(I_n - M^TM)J^T(y)M^T] \\ &= \text{tr } [MJ(y)(I_n - M^TM)J^T(y)M^T]. \end{aligned} \quad (46)$$

Following again the previous work on exact lumping, $J^T(y)$ can be decomposed into a linear combination of appropriate constant matrices A_k ($k = 1, 2, \dots, m$), i.e.

$$J^T(y) = \sum_{k=1}^m a_k(y) A_k. \quad (47)$$

The coefficients $a_k(y)$ are functions of y . Substituting eq. (47) into eq. (46) yields

$$\begin{aligned} Z_1(y) &= \text{tr } \left[M \sum_{k=1}^m a_k(y) A_k^T (I_n - M^TM) \right. \\ &\quad \left. \sum_{k'=1}^m a_{k'}(y) A_{k'} M^T \right] \\ &= \text{tr } \sum_{k,k'=1}^m a_k(y) a_{k'}(y) M A_k^T (I_n - M^TM) A_{k'} M^T. \end{aligned} \quad (48)$$

If y varies in a region Ω of the Y_n -composition space, the total error Z_1 can be denoted by the integration of $Z_1(y)$ over Ω :

$$\begin{aligned} Z_1 &= \int_{\Omega} Z_1(y) d\Omega \\ &= \text{tr } \sum_{k,k'=1}^m \int_{\Omega} a_k(y) a_{k'}(y) d\Omega \\ &\quad \times M A_k^T (I_n - M^TM) A_{k'} M^T \\ &= \text{tr } \sum_{k,k'=1}^m a_{kk'} M A_k^T (I_n - M^TM) A_{k'} M^T \end{aligned} \quad (49)$$

where

$$a_{kk'} = \int_{\Omega} a_k(y) a_{k'}(y) d\Omega. \quad (50)$$

The flexibility available in choosing Ω allows for tailoring the lumping as desired.

We need to determine a matrix M , which gives the smallest total error Z_1 . This problem can be de-

scribed as

$$\begin{aligned} \text{minimize } Z_1 &= \text{tr} \sum_{k,k'=1}^m a_{kk'} M A_k^T (I_n - M^T M) A_{k'} M^T \\ \text{subject to } M M^T &= I_{\hat{n}}. \end{aligned} \quad (51)$$

The constraint can be included by Lagrange's method of undetermined multipliers Let

$$\begin{aligned} Z'_1 &= \text{tr} \sum_{k,k'=1}^m a_{kk'} M A_k^T (I_n - M^T M) A_{k'} M^T \\ &+ \sum_{i,j=1}^{\hat{n}} \lambda_{ij} \left(\sum_{s=1}^n m_{is} m_{js} - \delta_{ij} \right), \end{aligned} \quad (52)$$

where λ_{ij} s are Lagrange multipliers, m_{kl} is the (k, l) -entry of M , and δ_{ij} is the Kronecker delta function.

In order to determine the matrix M we need to solve the following equations:

$$\begin{aligned} \partial Z'_1 / \partial M^T &= 0 \\ \partial Z'_1 / \partial \lambda_{ij} &= 0 \quad (\text{for all } i \text{ and } j). \end{aligned} \quad (53)$$

After some lengthy manipulation (Appendix A), we find that M must satisfy the following matrix equation:

$$\begin{aligned} (I_n - M^T M) \sum_{k,k'=1}^m a_{kk'} (A_k^T A_{k'} - A_k^T M^T M A_{k'}) \\ - A_k M^T M A_k^T M^T = 0. \end{aligned} \quad (54)$$

It is easy to demonstrate that, if a system is exactly lumpable, the corresponding matrix M of a fixed $J^T(\mathbf{y})$ -invariant subspace \mathcal{M} is a solution of eq. (54). Let us now consider uni- and/or bimolecular reaction systems. In this case, it has been proved in our previous paper that \mathcal{M} is simultaneously invariant under all A_k s, i.e.

$$A_k M^T = M^T P_k \quad (55)$$

where P_k is an \hat{n} -square constant matrix. Utilizing this relation one can readily prove the validity of eq. (54). The explicit dependence on A_k and $a_{kk'}$ in eq. (54) can be eliminated by substituting eqs (47) and (50) back into eq. (54) to yield an equation which contains $J(\mathbf{y})$ instead. Then the same conclusion can be obtained in the same way for other systems not easily decomposed to a linear combination of constant matrices. To save space we leave the demonstration to the reader.

Equation (54) is a nonlinear matrix equation, which is likely difficult to solve analytically. However, after expansion of eq. (54), we obtain $n \times \hat{n}$ nonlinear algebraic equations with the highest order 5 in the elements of M . The equations can be solved numerically by an iteration method, if one uses suitable initial values of M .

(2) *Constrained approximate lumping matrices.* Most probably the lumped model will satisfy some restrictions. For instance, some species may be left unlumped for practical purposes. The freedom in choosing Ω in eq. (50) also corresponds to a special

type of constraint, and this perspective will be discussed further in Section 4. The determination of the approximate lumping schemes under general constraints is an important problem. Constraints on the species can be included by specifying a part of the lumping matrix M and seeking to determine the remainder of it. This circumstance corresponds to the above situation of there being unlumped species, where the known part of M is just a submatrix with unit diagonal elements and zeros elsewhere. In this case the lumping matrix M can be represented as

$$M = \begin{pmatrix} M_G \\ M_D \end{pmatrix} \quad (56)$$

where M_G is given and also required to satisfy $M_G M_G^T = I$; M_D will be determined. Then we have

$$\begin{aligned} E_1(\mathbf{y}) &= (I_n - M^T M) J^T(\mathbf{y}) M^T \\ &= (I_n - M_G^T M_G - M_D^T M_D) \\ &\quad \times \sum_{k=1}^m a_k(\mathbf{y}) A_k (M_G^T M_G^T). \end{aligned} \quad (57)$$

Now the problem is expressed as

$$\begin{aligned} \text{minimize } Z_1 &= \text{tr} \sum_{k,k'=1}^m a_{kk'} \begin{pmatrix} M_G \\ M_D \end{pmatrix} A_k^T (I_n - M_G^T M_G \\ &\quad - M_D^T M_D) A_{k'} \begin{pmatrix} M_G^T \\ M_D^T \end{pmatrix} \\ \text{subject to } M M^T &= I_{\hat{n}}. \end{aligned} \quad (58)$$

Using the same approach as that for eq. (51), we find that M_D must satisfy the following matrix equation:

$$\begin{aligned} (I_n - M_G^T M_G - M_D^T M_D) \sum_{k,k'=1}^m a_{kk'} (A_k^T A_{k'} \\ - A_k^T M_G^T M_G A_{k'} - A_k M_G^T M_G A_k^T \\ - A_k^T M_D^T M_D A_{k'} - A_k M_D^T M_D A_k^T) M_D^T = 0. \end{aligned} \quad (59)$$

Equation (59) is almost the same as eq. (54) except for containing some constant matrices and eq. (59) can be obtained by substituting eq. (56) into eq. (54).

3D. The equation for determining M under minimization of E_2

Just like the minimization for E_1 , we define the error $Z_2(\mathbf{y})$ for given M and \mathbf{y} by the trace of matrix $E_2^T(\mathbf{y}) E_2(\mathbf{y})$, which is the sum of the squares of all the elements in $E_2(\mathbf{y})$. Since we have chosen $\bar{M} = M^T$ in Section 3B, then

$$\begin{aligned} Z_2(\mathbf{y}) &= \text{tr} [E_2^T(\mathbf{y}) E_2(\mathbf{y})] = \text{tr} [E_2(\mathbf{y}) E_2^T(\mathbf{y})] \\ &= \text{tr} \{ M [J(\mathbf{y}) - J(M^T M \mathbf{y})] [J^T(\mathbf{y}) \\ &\quad - J^T(M^T M \mathbf{y})] M^T \} = \text{tr} \{ M [J(\mathbf{y}) J^T(\mathbf{y}) \\ &\quad - J(\mathbf{y}) J^T(M^T M \mathbf{y}) - J(M^T M \mathbf{y}) J^T(\mathbf{y}) \\ &\quad + J(M^T M \mathbf{y}) J^T(M^T M \mathbf{y})] M^T \}. \end{aligned} \quad (60)$$

Let

$$\mathbf{z} = M^T M \mathbf{y}. \quad (61)$$

Like eq. (47) we have

$$J^T(M^T M y) = \sum_{k=1}^m a_k(z) A_k. \quad (62)$$

Utilizing eqs (47) and (62), eq. (60) becomes

$$Z_2(y) = \text{tr} \sum_{k,k'=1}^m [a_k(y) a_{k'}(y) - a_k(y) a_{k'}(z) - a_k(z) a_{k'}(y) + a_k(z) a_{k'}(z)] M A_k^T A_{k'} M^T. \quad (63)$$

If y varies in a region Ω of the Y_n -composition space, the total error Z_2 can be denoted by the integration of $Z_2(y)$ over Ω :

$$\begin{aligned} Z_2 &= \int_{\Omega} Z_2(y) d\Omega \\ &= \text{tr} \sum_{k,k'=1}^m \int_{\Omega} [a_k(y) a_{k'}(y) - a_k(y) a_{k'}(z) - a_k(z) a_{k'}(y) + a_k(z) a_{k'}(z)] d\Omega M A_k^T A_{k'} M^T \\ &= \text{tr} \sum_{k,k'=1}^m [a_{kk'} - b_{kk'}(M) - b_{k'k}(M) + c_{kk'}(M)] M A_k^T A_{k'} M^T \end{aligned} \quad (64)$$

where

$$a_{kk'} = \int_{\Omega} a_k(y) a_{k'}(y) d\Omega \quad (65)$$

$$b_{kk'}(M) = \int_{\Omega} a_k(y) a_{k'}(z) d\Omega \quad (66)$$

$$c_{kk'}(M) = \int_{\Omega} a_k(z) a_{k'}(z) d\Omega \quad (67)$$

and $b_{kk'}(M)$ and $c_{kk'}(M)$ are functions of M due to eq. (61). Let

$$\beta_{kk'}(M) = a_{kk'} - b_{kk'}(M) - b_{k'k}(M) + c_{kk'}(M). \quad (68)$$

Then we have

$$Z_2 = \text{tr} \sum_{k,k'=1}^m \beta_{kk'}(M) M A_k^T A_{k'} M^T. \quad (69)$$

Since $\beta_{kk'}(M)$ is a complicated function of M , it is very difficult to obtain the analytic solution of the equation arising from differentiation of Z_2 with respect to M . Therefore, we cannot obtain the corresponding equation to minimize Z_2 as eq. (54) or (59).

Thus far we have considered the determination of M from minimization of Z_1 and Z_2 separately, but in practice we seek a dual minimization of Z_1 and Z_2 to obtain M . Considering that Z_2 is a nonnegative number and the smaller the better, we can treat Z_2 as a parameter and choose an appropriate value of it, and then solve eqs (40), (54) [or (59)] and (69) simultaneously to determine M . We can choose the value of Z_2 as small as possible under the condition that the resultant Z_1 is acceptable. In this way the approximate lumping matrices M with orthonormal rows and minima E_1 and E_2 can be obtained.

4. DETERMINATION OF THE APPROXIMATE LUMPING MATRICES VALID IN A GIVEN REGION OF THE COMPOSITION Y_n -SPACE

In the foregoing section the equations to determine the approximate lumping matrices have been presented. For realistic problems the chosen initial compositions will usually constrain the system to some small region of composition space. Therefore the approximate lumping matrix validated for the whole composition space could give a quite large error for some given narrow region. Choosing a better lumping matrix in a given region becomes desirable, and multiple lumping matrices may be used to cover a large portion of composition space. Several lumping matrices of various dimensions n and quality might also exist in each region.

The derivations leading to eqs (54), (59) and (69) show that the determination of M follows the same procedure regardless of the size of the desired region. Equations (54), (59) and (69) contain the coefficients $a_{kk'}$ and $\beta_{kk'}(M)$ defined by eqs (65)–(68). These coefficients are evaluated in a given region Ω of the composition space. Thus different regions simply correspond to different values for the coefficients $a_{kk'}$ and $\beta_{kk'}(M)$. After the determination of $a_{kk'}$ and $\beta_{kk'}(M)$ in a given region, one can obtain the corresponding lumping matrices by solving eqs (40), (54) [or (59)] and (69) simultaneously.

4A. Determination of $a_{kk'}$ and $\beta_{kk'}(M)$ for the whole composition region

The whole composition region in realistic problems means that under the condition of the total quantity $c > 0$ of the reaction system being constant, any species can take on any value from 0 to the c . This is a rather special circumstance which can arise in certain applications (e.g. when y corresponds to a state population vector). Notice that in this case all y_k s are equivalent for the purpose of determining $a_{kk'}$ and $\beta_{kk'}(M)$ with

$$\sum_{i=1}^n y_i = c. \quad (70)$$

Then using eqs (65)–(67) we have

$$\begin{aligned} a_{kk'} &= \int_{\Omega} a_k(y) a_{k'}(y) d\Omega \\ &= \int_0^c \int_0^{c-y_1} \dots \int_0^{c-\sum_{i=1}^{n-1} y_i} a_k(y) \\ &\quad \times a_{k'}(y) dy_1 dy_2 \dots dy_n. \end{aligned} \quad (71)$$

$$\begin{aligned} b_{kk'}(M) &= \int_{\Omega} a_k(y) a_{k'}(z) d\Omega \\ &= \int_0^c \int_0^{c-y_1} \dots \int_0^{c-\sum_{i=1}^{n-1} y_i} a_k(y) \\ &\quad a_{k'}(M^T M y) dy_1 dy_2 \dots dy_n. \end{aligned} \quad (72)$$

$$\begin{aligned}
 c_{kk'}(M) &= \int_{\Omega} a_k(z) a_{k'}(z) d\Omega \\
 &= \int_0^c \int_0^{c-y_1} \cdots \int_0^{c-\sum_{i=1}^{n-1} y_i} a_k(M^T M y) \\
 &\quad a_{k'}(M^T M y) dy_1 dy_2 \cdots dy_n.
 \end{aligned} \quad (73)$$

Returning to the uni- and/or bimolecular reaction systems, we proved that the transpose of the Jacobian matrix $J^T(y)$ can be expressed as (Li and Rabitz, 1989)

$$J^T(y) = A_0 + \sum_{k=1}^n y_k A_k. \quad (74)$$

Then

$$\begin{aligned}
 a_{00} &= \int_0^c \int_0^{c-y_1} \cdots \int_0^{c-\sum_{i=1}^{n-1} y_i} dy_1 dy_2 \cdots dy_n \\
 &= \frac{c^n}{n!}.
 \end{aligned} \quad (75)$$

Using the equivalence of the y_k s we can change the order of y_k s such that $y_k = y_1$ and $y_{k'} = y_2$. Then we have

$$\begin{aligned}
 a_{0k} &= a_{k0} = a_{01} \\
 &= \int_0^c \int_0^{c-y_1} \cdots \int_0^{c-\sum_{i=1}^{n-1} y_i} y_1 dy_1 dy_2 \cdots dy_n \\
 &= \frac{c^{n+1}}{(n+1)!}.
 \end{aligned} \quad (76)$$

$$\begin{aligned}
 a_{kk} &= a_{11} = \int_0^c \int_0^{c-y_1} \cdots \int_0^{c-\sum_{i=1}^{n-1} y_i} y_1^2 dy_1 dy_2 \\
 &\quad \cdots dy_n = \frac{2c^{n+2}}{(n+2)!}.
 \end{aligned} \quad (77)$$

$$\begin{aligned}
 a_{kk'} &= a_{12} \\
 &= \int_0^c \int_0^{c-y_1} \cdots \int_0^{c-\sum_{i=1}^{n-1} y_i} y_1 y_2 dy_1 dy_2 \cdots dy_n \\
 &= \frac{c^{n+2}}{(n+2)!} \quad (k \neq k').
 \end{aligned} \quad (78)$$

In the same way, we can determine $b_{kk'}(M)$ and $c_{kk'}(M)$:

$$b_{00}(M) = \int_{\Omega} d\Omega = \frac{c^n}{n!} \quad (79)$$

$$\begin{aligned}
 b_{0k} &= \int_{\Omega} z_k d\Omega \\
 &= \int_{\Omega} \sum_{i=1}^n h_{ki} y_i d\Omega = \sum_{i=1}^n h_{ki} \int_{\Omega} y_i d\Omega \\
 &= \sum_{i=1}^n h_{ki} \frac{c^{n+1}}{(n+1)!},
 \end{aligned} \quad (80)$$

where

$$h_{ki} = \sum_{s=1}^n m_{sk} m_{si} \quad (81)$$

and m_{ij} is the (i, j) -entry of M .

$$\begin{aligned}
 b_{k0} &= \int_{\Omega} y_k d\Omega \\
 &= \frac{c^{n+1}}{(n+1)!}.
 \end{aligned} \quad (82)$$

$$\begin{aligned}
 b_{kk'} &= \int_{\Omega} y_k z_{k'} d\Omega \\
 &= \sum_{i=1}^n h_{ki} \int_{\Omega} y_i y_{k'} d\Omega + h_{k'k} \int_{\Omega} y_k^2 d\Omega \\
 &= \sum_{i=1}^n h_{ki} \frac{c^{n+2}}{(n+2)!} + h_{k'k} \frac{2c^{n+2}}{(n+2)!} \\
 &= \left(h_{k'k} + \sum_{i=1}^n h_{ki} \right) \frac{c^{n+2}}{(n+2)!}.
 \end{aligned} \quad (83)$$

Similarly, we also have

$$c_{00}(M) = \int_{\Omega} d\Omega = \frac{c^n}{n!} \quad (84)$$

$$\begin{aligned}
 c_{0k} &= c_{k0} = \int_{\Omega} z_k d\Omega \\
 &= \sum_{i=1}^n h_{ki} \frac{c^{n+1}}{(n+1)!},
 \end{aligned} \quad (85)$$

$$\begin{aligned}
 c_{kk'} &= \int_{\Omega} z_k z_{k'} d\Omega \\
 &= \left(\sum_{i,j=1}^n h_{ki} h_{k'j} + \sum_{i=1}^n h_{ki} h_{k'i} \right) \frac{c^{n+2}}{(n+2)!}.
 \end{aligned} \quad (86)$$

Without any loss of generality it is convenient to normalize the composition unit such that

$$c = 1. \quad (87)$$

Then the coefficients $a_{kk'}$, $b_{kk'}(M)$ and $c_{kk'}(M)$ have simple values:

$$\begin{aligned}
 a_{00} &= \frac{1}{n!} \\
 a_{0k} &= a_{k0} = \frac{1}{(n+1)!} \\
 a_{kk} &= \frac{2}{(n+2)!} \\
 a_{kk'} &= \frac{1}{(n+2)!}.
 \end{aligned} \quad (88)$$

Multiplying all $a_{kk'}$ by the same constant will not affect the solutions of eqs (54), (59) and (69). Hence, we can use another set of $a_{kk'}$:

$$\begin{aligned}
 a_{00} &= (n+2)(n+1) \\
 a_{0k} &= a_{k0} = n+2 \\
 a_{kk} &= 2 \\
 a_{kk'} &= 1.
 \end{aligned} \quad (89)$$

Following the same procedure we have

$$\begin{aligned} b_{00} &= (n+2)(n+1) \\ b_{0k} &= \left(\sum_{i=1}^n h_{ki} \right) (n+2) \\ b_{k0} &= n+2 \\ b_{kk'} &= h_{k'k} + \sum_{i=1}^n h_{k'i} \end{aligned} \quad (90)$$

$$\begin{aligned} c_{00} &= (n+2)(n+1) \\ c_{0k} &= c_{k0} = \left(\sum_{i=1}^n h_{ki} \right) (n+2) \\ c_{kk'} &= \sum_{i,j=1}^n h_{ki} h_{k'j} + \sum_{i=1}^n h_{ki} h_{k'i} \end{aligned} \quad (91)$$

Substituting above equations into eq. (68) yields

$$\begin{aligned} \beta_{00} &= 0 \\ \beta_{0k} &= \beta_{k0} = 0 \\ \beta_{kk} &= 2 - 2 \left(h_{kk} + \sum_{i=1}^n h_{ki} \right) + \sum_{i,j=1}^n h_{ki} h_{kj} \\ &\quad + \sum_{i=1}^n h_{ki}^2 \\ \beta_{kk'} &= \beta_{k'k} = 1 - (h_{k'k} + h_{kk'}) \\ &\quad - \sum_{i=1}^n (h_{ki} + h_{k'i} - h_{ki} h_{k'i}) + \sum_{i,j=1}^n h_{ki} h_{k'j} \end{aligned} \quad (92)$$

4B. Determination of $a_{kk'}$ and $\beta_{kk'}(M)$ for a reaction path

Let us consider Ω as a reaction path in composition space. In this case

$$d\Omega = ds \quad (93)$$

$$a_{kk'} = \int_0^{s_f} a_k(y) a_{k'}(y) ds \quad (94)$$

where s is the length of the reaction path and s_f represents the final value of the given reaction path in the composition space. Since

$$ds = \left(\sum_{i=1}^n dy_i^2 \right)^{1/2} \quad (95)$$

we can determine $a_{kk'}$ numerically by either one of the following two means for a given initial $y(0)$:

$$\begin{aligned} a_{kk'} &= \int_0^\infty a_k(y) a_{k'}(y) \left[\sum_{i=1}^n \left(\frac{dy_i}{dt} \right)^2 \right]^{1/2} dt \\ &= \int_0^\infty a_k(y) a_{k'}(y) \left[\sum_{i=1}^n f_i^2(y) \right]^{1/2} dt \end{aligned} \quad (96)$$

or

$$\begin{aligned} a_{kk'} &= \int_{y_{1i}}^{y_{1f}} a_k(y) a_{k'}(y) \left[\sum_{i=1}^n \left(\frac{dy_i}{dy_1} \right)^2 \right]^{1/2} dy_1 \\ &= \int_{y_{1i}}^{y_{1f}} a_k(y) a_{k'}(y) \left\{ \sum_{i=1}^n [f_i(y)/f_1(y)]^2 \right\}^{1/2} dy_1 \end{aligned} \quad (97)$$

where y_{1i} and y_{1f} are the initial and final values of y_1 ,

respectively. Since $a_k(y)$, $f_i(y)$ and initial $y(0)$ have been given, one can obtain $a_{kk'}$ numerically by these equations. Similarly, one can determine the corresponding $b_{kk'}(M)$ and $c_{kk'}(M)$, and consequently $\beta_{kk'}(M)$ by these equations through the replacement of $a_k(y)a_{k'}(y)$ by $a_k(y)a_{k'}(z)$ and $a_k(z)a_{k'}(z)$, respectively.

4C. Determination of $a_{kk'}$ and $\beta_{kk'}(M)$ for a given region of the composition space

For most realistic problems the initial composition is constrained to be chosen from a given region. Suppose the initial composition contains only l species taking values in the following regions:

$$\begin{aligned} y_{1i}^0 &\leq y_1^0 \leq y_{1f}^0 \\ y_{2i}^0 &\leq y_2^0 \leq y_{2f}^0 \\ &\vdots \\ y_{li}^0 &\leq y_l^0 \leq y_{lf}^0 \end{aligned} \quad (98)$$

where y_{li}^0 and y_{lf}^0 are the boundary values of the initial concentration for species y_l . In this case the $a_{kk'}$ equals the sum of those $a_{kk'}$ s of a reaction path with the initial values located in the above region and can be calculated numerically by the following equation:

$$a_{kk'} = \int_{y_{1i}^0}^{y_{1f}^0} \int_{y_{2i}^0}^{y_{2f}^0} \dots \int_{y_{li}^0}^{y_{lf}^0} a_{kk'}(y^0) dy_1^0 dy_2^0 \dots dy_l^0 \quad (99)$$

where $a_{kk'}(y^0)$ can be determined by eq. (96) or (97). Following the same procedure one can determine $\beta_{kk'}(M)$. In this fashion we can obtain $a_{kk'}$ and $\beta_{kk'}(M)$ that are associated with a volume in composition space.

5. THE CHOICE OF INITIAL VALUES FOR THE EQUATIONS TO DETERMINE M

In Section 3 we obtained the equations for determining M . However, eqs (54) and (59) give all the minima, maxima and other stationary points of the total error Z_1 . The particular type of solutions we obtain by an iteration method will depend on the chosen initial values of M . In most cases we are only interested in the global minima, but there is no easy way to determine them. In some cases solutions at local minima may suffice, since choices of acceptable M can also be guided by additional criteria besides minimization of Z_1 and Z_2 . When the dimension of M is high, the number of solutions for the equations to determine M becomes very large, as does the region of the initial values of M we can choose from. It is thus impossible to randomly search the entire region. Therefore we must develop a logical approach for choosing the initial values.

We know from previous work that, if a system is exactly lumpable, the solutions of eqs (54), (59) and (69) are the matrix representations of the simultaneously invariant subspaces of all A_k s. When a system does not have such a subspace with a given dimension, the corresponding subspaces of the global minimum solutions of these equations should be very

close to the invariant subspaces of all the A_k s. Certainly, the solutions are not expected to be equally close to each A_k -invariant subspace, because the coefficients $a_k(y)$ s give the A_k s different weights. This property suggests an approach for choosing the suitable initial values. If we can find a group of M such that the corresponding subspace of each M has very high degrees of coincidence with the invariant subspaces of all the A_k s, these M will definitely give small Z_1 and one of them will give small Z_2 . Then these choices can be taken as initial values of M to minimize Z_1 and Z_2 .

In order to achieve this task, the procedure to determine these initial choices for M consists of two steps. First, we determine the groups of m \hat{n} -dimensional subspaces, each one of which is invariant to one A_k ($k = 1, 2, \dots, m$). These groups will have the highest sums of degrees of coincidence between each pair of invariant subspaces compared to other groups. This means that the invariant subspaces of A_k s in these groups are the closest to one another. Second, we determine the \hat{n} -dimensional subspace \mathcal{M} , which has the highest sum of degrees of coincidence with each invariant subspace in one of these closest groups. Then \mathcal{M} is the subspace which has the highest degrees of coincidence with the invariant subspaces of all the A_k s. Therefore, the matrix representations of \mathcal{M} can be used as the initial values of M .

As shown in our previous paper (Li and Rabitz, 1989), the invariant subspaces of A_k can be obtained through its Jordan canonical form. If the number of the invariant subspaces for each A_k is finite, all the groups of the invariant subspaces, each one of which comes from one A_k , can be examined. When the number is infinite, we are not able to examine all of them. Therefore, some good initial estimates of M may possibly be lost. Nevertheless, this approach will supply some suitable initial values of M .

We must now establish how to determine the degree of coincidence of two subspaces. Here we simply give the approach; the details of it can be found in Appendix B. Suppose $\mathcal{M}(r)$ and $\mathcal{M}(r')$ are r - and r' -dimensional subspaces, respectively. We choose corresponding r and r' orthonormal vectors as their bases. Let the $n \times r$ and $n \times r'$ matrices $Y(r)$ and $Y(r')$ be the matrix representations of the two subspaces with $r' \leq r$. The degree of coincidence d_c of the two subspaces is defined as follows:

$$d_c = \frac{1}{r'} \operatorname{tr} [Y(r')^T Y(r) Y(r)^T Y(r')]. \quad (100)$$

When one of the two subspace is contained within the other one, $d_c = 1$. When the two subspaces are orthogonal to each other, $d_c = 0$. In other cases, $0 < d_c < 1$. It may also be proved that d_c is independent of the choice of the orthonormal bases of $\mathcal{M}(r)$ and $\mathcal{M}(r')$.

Using the definition of degree of coincidence between two subspaces we can determine d_c for any two subspaces with dimension \hat{n} , each of which is invariant to different A_k . Then we can find the closest

groups of m A_k -invariant subspaces with the same dimension \hat{n} , which have the largest sums of d_c . It is not necessary that each A_k -invariant subspace has dimension \hat{n} . The A_k -invariant subspace can have dimension larger than \hat{n} , if any subspace of it is also A_k -invariant.

Suppose we have found one of the closest groups of the invariant subspaces of the A_k s, whose corresponding matrix representations are $Y_1(r_1)$, $Y_2(r_2)$, \dots , $Y_m(r_m)$ with dimension r_k equal to or larger than \hat{n} . The columns of each $Y_k(r_k)$ are orthonormal. Now we need to determine the initial value of an $\hat{n} \times n$ matrix M . The best estimate is the matrix representation of the subspace which has the largest sum of the degrees of coincidence with all $Y_k(r_k)$ s. Suppose the transpose of the best initial estimate of the solution is denoted by an $n \times \hat{n}$ matrix M^T , which also has orthonormal columns, then the sum of degrees of coincidence S between M^T and all $Y_k(r_k)$ s can be expressed as

$$S = \max_{MM^T = I_{\hat{n}}} \operatorname{tr} M \left[\frac{1}{\hat{n}} \sum_{k=1}^m Y_k(r_k) Y_k^T(r_k) \right] M^T \\ = \max_{MM^T = I_{\hat{n}}} \operatorname{tr} M Y M^T \quad (101)$$

where

$$Y = \frac{1}{\hat{n}} \sum_{k=1}^m Y_k(r_k) Y_k^T(r_k). \quad (102)$$

The solution to the problem in eq. (101) has been obtained (Bellman, 1970). Let the λ_i s represent the eigenvalues of Y and

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n.$$

The corresponding eigenvector matrix is R and the first k columns of R are denoted by $R_{(k)}$. Then we have

$$\max_{MM^T = I_{\hat{n}}} \operatorname{tr} M Y M^T = \sum_{i=1}^{\hat{n}} \lambda_i \quad (103)$$

$$M^T = R_{(\hat{n})}. \quad (104)$$

Therefore, when we have determined one of the closest groups of invariant subspaces of the A_k s, we can get Y and its eigenvectors, which are arranged by nonincreasing order of their eigenvalues. Then the first \hat{n} eigenvectors are a good initial estimate of the solution M^T . Notice that Y is a symmetric matrix and it has full eigenvectors. Any linear combination of the eigenvectors corresponding to a multiple eigenvalue is still an eigenvector of Y . Therefore, when Y has multiple eigenvalues, sometimes the solution is not unique. All the combinations of the eigenvectors with the same largest sums of corresponding eigenvalues are solutions. Since this approach only supplies good estimates of M and the global minimum solutions in our problem usually are not unique, the first several closest groups should be used to construct initial values of M .

If we need to determine a constrained approximate lumping matrix, then in this case eq. (101) becomes

$$\begin{aligned}
 S &= \max_{MM^T = I_2} \text{tr} \begin{pmatrix} M_G \\ M_D \end{pmatrix} Y (M_G^T M_D^T) \\
 &= \max_{MM^T = I_2} (\text{tr } M_G Y M_G^T + \text{tr } M_D Y M_D^T) \\
 &= S_G + S_D.
 \end{aligned} \quad (105)$$

The second term S_D on the right-hand side of eq. (105) is just the same as the S of unconstrained approximate lumping. Therefore, after the determination of the closest groups we compute the corresponding values of the first term S_G for the given Y s and then we choose the solutions with the largest total S as the initial estimates of constrained approximate lumping matrices.

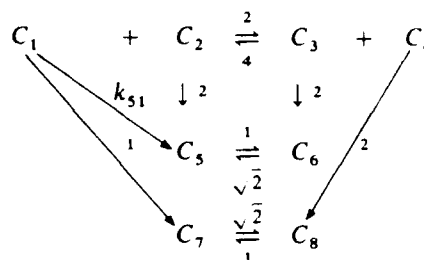
In most cases, the number of the invariant subspaces of A_k is infinite. Sometimes, we cannot examine all the groups of the A_k -invariant subspaces with dimensions from 1 to $n - 1$. Hence, we may fail to find suitable initial values of M from the closest groups owing to incomplete examination or there only being available lower dimensional A_k -invariant subspaces. In order to treat this problem we can extend the above method in two ways. We can use the sums of the lower-dimension solutions obtained from different closest groups to give the estimates of the higher-dimension solutions. The only thing we need to do is to orthonormalize these solutions so that the initial estimate of M satisfies the restriction of $MM^T = I_n$. Second, we can use the "expanded" invariant subspace corresponding to eigenvalues which are almost equal. In this case, any subspace in the expanded one is almost invariant to its original matrix. Therefore, we can determine M with higher dimensions. This approach will be illustrated by the following examples.

$$J^T(y) = \begin{bmatrix} -2y_2 - 1 - k_{s1} & -2y_2 & 2y_2 & 2y_2 & k_{s1} & 0 & 1 & 0 \\ -2y_1 & -2(1 + y_1) & 2y_1 & 2y_1 & 2 & 0 & 0 & 0 \\ 4y_4 & 4y_4 & -2(1 + 2y_4) & -4y_4 & 0 & 2 & 0 & 0 \\ 4y_3 & 4y_3 & -4y_3 & -2(1 + 2y_3) & 0 & 0 & 0 & 2 \\ & 0 & & & -1 & 1 & 0 & 0 \\ & & & & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ 0 & 0 & & & 0 & 0 & 1 & -1 \end{bmatrix}$$

6. EXAMPLES

The method proposed in this paper will be illustrated by the following reaction scheme, where the C_i s

are species and the numbers are unitless rate constants.



This is a modification of an example used in our previous paper where $k_{s1} = 1$ admitted some exact lumping solutions. By changing the rate constant k_{s1} to 0.9 (example 1) and 0.1 (example 2) the system contains some exact and approximate lumping schemes. The focus here should be on the approximate lumping schemes, since in real problems the presence of nontrivial exact lumping is not likely. If exact lumping schemes exist, they should be obtained by the present approach corresponding to the special case $Z_1 = Z_2 = 0$.

Letting y_i represent the concentration of C_i , it is easy to write out the kinetic equations and the transpose of the corresponding Jacobian matrix $J^T(y)$.

$$\begin{aligned}
 dy_1/dt &= -(1 + k_{s1})y_1 - 2y_1y_2 + 4y_3y_4 \\
 dy_2/dt &= -2y_2 - 2y_1y_2 + 4y_3y_4 \\
 dy_3/dt &= -2y_3 - 4y_3y_4 + 2y_1y_2 \\
 dy_4/dt &= -2y_4 - 4y_3y_4 + 2y_1y_2 \\
 dy_5/dt &= -y_5 + k_{s1}y_1 + 2y_2 + \sqrt{2}y_6 \\
 dy_6/dt &= -\sqrt{2}y_6 + 2y_3 + y_5 \\
 dy_7/dt &= -\sqrt{2}y_7 + y_1 + y_8 \\
 dy_8/dt &= -y_8 + 2y_4 + \sqrt{2}y_7
 \end{aligned} \quad (106)$$

$J^T(y)$ can be represented as

$$J^T(y) = A_0 + \sum_{k=1}^4 y_k A_k \quad (107)$$

where

$$A_0 = \begin{pmatrix} -1 - k_{s1} & 0 & 0 & 0 & k_{s1} & 0 & 1 & 0 \\ 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2 \\ & & & & -1 & 1 & 0 & 0 \\ & & & & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ 0 & 0 & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ 0 & 0 & & & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$\lambda_i = \begin{matrix} & -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_3} = \begin{pmatrix} 0.0000 & 0.7071 & -0.4083 & 0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.8165 & 0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.8660 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.0000 & 0.7071 & 0.4083 & -0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

$$\lambda_i = \begin{matrix} & -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_4} = \begin{pmatrix} 0.0000 & 0.7071 & -0.4083 & 0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.8165 & 0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.0000 & 0.7071 & 0.4083 & -0.2887 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.8660 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

Notice that any linear combination of the eigenvectors corresponding to a multiple eigenvalue of A_k is still an eigenvector of it. The subspace spanned by all the eigenvectors corresponding to an eigenvalue of A_k is a root subspace and any subspace of the root one is A_k -invariant. Hence, there are an infinite number of invariant subspaces for each A_k and we cannot examine all of them. However, using the property of the root subspace mentioned above, we can determine the closest groups of the root subspaces, each one of which comes from an X_{A_k} ($k = 0, 1, \dots, 4$). These closest groups can be used to choose some initial values of M with \hat{n} not larger than the smallest dimension of these root subspaces.

(1) *Unconstrained lumping matrices.* Let $Y_k(i)$ represent the i th submatrix of X_{A_k} corresponding to the i th distinct eigenvalue listed above each matrix X_{A_k} . For example, the first column of X_{A_0} is $Y_0(1)$, columns 2–4 of X_{A_0} is $Y_0(2)$, etc. The columns of $Y_k(i)$ span a root subspace. For convenience, the columns are taken as being orthonormal. Since -1.9 and -2.0 are very close eigenvalues of A_0 , the first four eigenvectors of A_0 are considered as spanning an expanded root subspace. Thus A_0 is approximately regarded as having three root subspaces with dimensions 4, 2 and 2. Each of the other A_k s has two root subspaces with dimensions 1 and 7.

Arbitrarily choosing one $Y_k(i_k)$ from each X_{A_k} one can compose a five-member group. Then using eq. (100) the degree of coincidence $d_c(k, k')$ for any pair of $Y_k(i_k)$ and $Y_{k'}(i_{k'})$ can be computed. Let D_c represent the sum of all the $d_c(k, k')$ in this group, i.e.

$$D_c = \sum_{\substack{k, k'=0 \\ k < k'}}^4 d_c(k, k'). \quad (108)$$

Comparing all the resultant D_c will yield the closest groups of the root subspaces for all A_k s. Notice that in each group there are 10 pairs of $Y_k(i_k)$ and $Y_{k'}(i_{k'})$ and the largest value of d_c is 1. Therefore, the maximum value of D_c is 10. The first several closest groups with the largest D_c obtained by eqs (100) and (108) are given in Table 1.

After the determination of the closest groups of the root subspaces for all A_k , we can use eqs (101)–(104) to find the initial estimates of M with different \hat{n} . The first closest group of the root subspaces for all A_k with $D_c = 9.9348$ consists of $Y_0(3)$ and $Y_k(2)$ ($k = 1-4$). $Y_0(3)$ has two columns and other $Y_k(2)$ s have seven columns. Therefore, this group can be only used to give the initial estimates of M with $\hat{n} = 1$ and 2. The corresponding matrices Y for $\hat{n} = 1$ and 2 can be obtained by eq. (102). Let $Y(i)$ represent the matrix Y for the i th group in Table 1. Then one can computationally determine the eigenvalues λ_i and eigenvector matrix R for the symmetric matrix Y . Similarly, we

Table 1. Sum D_c of degrees of coincidence for the largest groups

No.	$Y_0(i)$	$Y_1(i)$	$Y_2(i)$	$Y_3(i)$	$Y_4(i)$	D_c
1	$Y_0(3)$	$Y_1(2)$	$Y_2(2)$	$Y_3(2)$	$Y_4(2)$	9.9348
2	$Y_0(1)$	$Y_1(2)$	$Y_2(2)$	$Y_3(2)$	$Y_4(2)$	9.0000
3	$Y_0(3)$	$Y_1(2)$	$Y_2(2)$	$Y_3(1)$	$Y_4(2)$	8.5076
4	$Y_0(1)$	$Y_1(2)$	$Y_2(2)$	$Y_3(1)$	$Y_4(2)$	8.5000

use $R(i)$ to represent the eigenvector matrix R of $Y(i)$. From eq. (102) we know that the difference between the two Y s for $\hat{n} = 1$ and 2 is a constant factor. Therefore, the corresponding $R(1)$ s are the same. The resultant $R(1)$ and the corresponding eigenvalues for

examine the two matrices $R(1)$ and $R(2)$ and their corresponding eigenvalues.

For $R(1)$ and $\hat{n} = 1$, the largest value of S is $\lambda_1 = 5$ and the first column of $R(1)$ is the best initial estimate of M^T , which is simply the trivial exact lumping scheme:

$$M = (0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536 \quad 0.3536).$$

different \hat{n} are as follows. The eigenvectors in $R(1)$ are arranged by nonincreasing order of their corresponding eigenvalues.

The second column of $R(1)$ has $S = \lambda_2 = 4.97$ almost equal to 5. It is also a quite good estimate of M^T with $\hat{n} = 1$. When $\hat{n} = 2$, the largest value of S is

$$R(1) = \begin{pmatrix} 0.3536 & 0.1407 & 0.1705 & -0.6837 & 0.3218 & 0.0000 & 0.0000 & 0.5053 \\ 0.3536 & -0.2668 & -0.6669 & 0.3414 & 0.0359 & 0.0000 & 0.0000 & 0.4907 \\ 0.3536 & -0.4142 & -0.2991 & -0.5027 & -0.3085 & 0.0000 & 0.0000 & -0.5152 \\ 0.3536 & 0.3599 & -0.1973 & 0.1604 & 0.6662 & 0.0000 & 0.0000 & -0.4874 \\ 0.3536 & -0.3405 & 0.4427 & 0.2426 & 0.0601 & -0.2695 & 0.6537 & -0.0123 \\ 0.3536 & -0.3405 & 0.4427 & 0.2426 & 0.0601 & 0.2695 & -0.6537 & -0.0123 \\ 0.3536 & 0.4336 & 0.0538 & 0.0997 & -0.4177 & 0.6537 & 0.2695 & 0.0156 \\ 0.3536 & 0.4336 & 0.0538 & 0.0997 & -0.4177 & -0.6537 & -0.2695 & 0.0156 \end{pmatrix}$$

\hat{n}	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
1	5.00	4.97	4.00	4.00	4.00	4.00	4.00	0.03
2	2.50	2.49	2.00	2.00	2.00	2.00	2.00	0.02

For the second closest group with $D_c = 9.0$, The resultant $R(2)$ of $Y(2)$ with different \hat{n} are also the same. In this case $Y_0(1)$ has the smallest number of columns 4. Then we can only use $R(2)$ to determine the initial estimates of M with \hat{n} from 1 to 4.

$\lambda_1 + \lambda_2 = 4.99$. Therefore, the first two columns of $R(1)$ can be used to construct the initial estimate of M^T with $\hat{n} = 2$. Other combinations of any two columns in $R(1)$ are not suitable for the initial estimates

$$R(2) = \begin{pmatrix} 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 \\ 0.5000 & -0.5000 & -0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 \\ 0.5000 & -0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.5000 \\ 0.5000 & 0.5000 & -0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.5000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 \end{pmatrix}$$

\hat{n}	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
1	5.00	5.00	5.00	4.00	4.00	4.00	4.00	1.00
2	2.50	2.50	2.50	2.00	2.00	2.00	2.00	0.50
3	1.67	1.67	1.67	1.33	1.33	1.33	1.33	0.33
4	1.25	1.25	1.25	1.00	1.00	1.00	1.00	0.25

Notice that in eq. (101) S is the sum of the degrees of coincidence between M^T and all $Y_k(r_k)$. Since there are only five $Y_k(r_k)$ in this example, the maximum value of S is 5, which corresponds to $Z_1 = 0$. Therefore, $5 - S$ can be applied as a reference value of Z_1 . Now let us

of M^T with $\hat{n} = 2$, because the corresponding S is considerably smaller than 5.

There are multiple eigenvalues in $R(2)$ and hence the solutions are not unique for different \hat{n} . When $\hat{n} = 1$ any one of the first three columns or any linear

(2) *Constrained lumping matrices.* Now let us consider the initial estimates of M under some con-

Using the approach presented in our previous paper on exact lumping, one can find that under this constraint the exact lumping matrices have \hat{n} higher than

4. For example the exact lumping matrix with $\hat{n} = 5$ is as follows:

In this case the system can be only approximately lumped by the lumping matrices, which contain M_G and have \hat{n} less than 5. Thus we need to determine the

other part M_D . Utilizing the resultant estimates of one-dimensional unconstrained lumping schemes from $R(1)$, $R(2)$ and M_G gives the following initial estimates of M with $n = 2$:

[illegible]

$$M_2 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & -0.5000 & -0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & -0.5000 & 0.5000 & -0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_5 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_6 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.7071 & 0.0000 & 0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

Notice that after orthonormalization M_1 becomes M_2 . All these matrices have $S_G = 2.00$, $S_D = 2.50$ and $S = 4.50$. They were used as initial values of M with $\hat{n} = 2$ and the best result was obtained by using M_2 as the initial value of M .

Similarly using the estimates of unconstrained two-dimensional lumping matrices obtained above and M_G we can construct the initial estimates of M with $\hat{n} = 3$:

$$M_7 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & -0.5000 & -0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_8 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & -0.5000 & 0.5000 & -0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_9 = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M_{10} = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.7071 & 0.0000 & 0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

After orthonormalization M_{10} becomes M_{11} :

$$M_{11} = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.8165 & 0.4083 & 0.4083 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

All these matrices have $S_G = 1.33$, $S_D = 3.33$ and $S = 4.66$. We cannot distinguish which is better. The best result was obtained by using M_9 as the initial value of M .

For $\hat{n} = 4$ we have the following initial estimate of M with $S_G = 1.00$, $S_D = 3.75$ and $S = 4.75$:

$$M_{12} = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5000 & 0.5000 & 0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & -0.5000 & -0.5000 & 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & -0.5000 & 0.5000 & -0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

Using these matrices as initial estimates of M with different \hat{n} and taking values of a_{kk} and $\beta_{kk}(M)$ from eqs (89) and (92) we solved eqs (40), (59) and (69) simultaneously by IMSL nonlinear equation system solver ZSCNT. The value of Z_2 was chosen in such a way that both Z_1 and Z_2 are acceptable. Notice that eqs (40) and (59) contain $(n + \hat{n}) \times \hat{n}$ nonlinear algebraic equations and eq. (69) has only one. In order to

force the solution to satisfy eq. (69) we can multiply this equation by a constant to increase its weight in this simultaneous nonlinear algebraic equation system. The resultant approximate lumping matrices validated in the whole composition region with different \hat{n} and the corresponding Z_1 and Z_2 are given

below:

$$\begin{array}{ccccc} \hat{n} & 2 & 3 & 4 & \\ Z_1 & 1.67 \times 10^{-2} & 1.65 \times 10^{-2} & 6.84 \times 10^{-3} & \\ Z_2 & 3.48 \times 10^{-4} & 2.91 \times 10^{-4} & 9.61 \times 10^{-4} & \end{array}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4843 & 0.5101 & 0.5026 & 0.5026 & -0.0012 & 0.0040 & -0.0072 & 0.0044 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4839 & 0.5098 & 0.5029 & 0.5029 & -0.0013 & 0.0040 & -0.0073 & 0.0047 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5211 & 0.4721 & 0.4913 & 0.5139 & -0.0052 & 0.0051 & -0.0030 & 0.0031 \\ 0.0338 & -0.0186 & 0.7137 & -0.6994 & -0.0002 & 0.0002 & -0.0001 & 0.0001 \\ -0.6534 & 0.7188 & 0.0484 & -0.0033 & 0.0055 & -0.0076 & 0.0049 & -0.0028 \end{pmatrix}$$

Choosing $\bar{M} = M^T$, we obtain the lumped kinetic equations from eq. (16). These lumped systems do not follow uni- and/or bimolecular reaction schemes, but this causes no real difficulty for practical purposes.

Lumped kinetic equations with $\hat{n} = 2$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.781619\hat{y}_2 \\ d\hat{y}_2/dt &= 1.325483 \times 10^{-3}\hat{y}_1 - 1.821451\hat{y}_2 \\ &\quad - 2.415180 \times 10^{-2}\hat{y}_2^2 \end{aligned} \quad (109)$$

Lumped kinetic equations with $\hat{n} = 3$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.975335\hat{y}_2 \\ d\hat{y}_2/dt &= 1.383473 \times 10^{-3}\hat{y}_1 - 1.973376\hat{y}_2 \\ &\quad - 9.616696 \times 10^{-4}\hat{y}_3 - 6.340748 \\ &\quad \times 10^{-3}\hat{y}_2^2 + 2.446022 \times 10^{-2}\hat{y}_3^2 \end{aligned} \quad (110)$$

$$d\hat{y}_3/dt = -2.000018\hat{y}_3$$

Lumped kinetic equations with $\hat{n} = 4$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.972365\hat{y}_2 + 2.77395 \times 10^{-2}\hat{y}_3 \\ &\quad + 0.105221\hat{y}_4 \\ d\hat{y}_2/dt &= -8.677774 \times 10^{-4}\hat{y}_1 - 1.973573\hat{y}_2 \\ &\quad + 4.623911 \times 10^{-3}\hat{y}_3 - 3.776060 \\ &\quad \times 10^{-2}\hat{y}_4 - 9.661564 \times 10^{-4}\hat{y}_2\hat{y}_3 \\ &\quad + 1.783237 \times 10^{-5}\hat{y}_2\hat{y}_4 \\ &\quad + 2.646131 \times 10^{-3}\hat{y}_3\hat{y}_4 - 5.257743 \\ &\quad \times 10^{-3}\hat{y}_2^2 + 2.410400 \times 10^{-2}\hat{y}_3^2 \\ &\quad - 1.203495 \times 10^{-2}\hat{y}_4^2 \\ d\hat{y}_3/dt &= -3.935029 \times 10^{-5}\hat{y}_1 + 1.755288 \\ &\quad \times 10^{-3}\hat{y}_2 - 1.999703\hat{y}_3 \\ &\quad - 2.376265 \times 10^{-3}\hat{y}_4 + 8.071963 \\ &\quad \times 10^{-5}\hat{y}_2\hat{y}_3 - 1.490951 \times 10^{-6}\hat{y}_2\hat{y}_4 \end{aligned}$$

$$\begin{aligned} &-2.212411 \times 10^{-4}\hat{y}_3\hat{y}_4 + 5.232053 \\ &\times 10^{-4}\hat{y}_2^2 - 2.015318 \times 10^{-3}\hat{y}_3^2 \\ &+ 1.006234 \times 10^{-3}\hat{y}_4^2 \end{aligned} \quad (111)$$

$$\begin{aligned} d\hat{y}_4/dt &= 1.091453 \times 10^{-3}\hat{y}_1 - 1.012915 \\ &\times 10^{-2}\hat{y}_2 - 9.035070 \times 10^{-3}\hat{y}_3 \\ &- 1.951685\hat{y}_4 - 1.574803 \times 10^{-3}\hat{y}_2\hat{y}_3 \\ &+ 2.906617 \times 10^{-5}\hat{y}_2\hat{y}_4 + 4.313106 \\ &\times 10^{-3}\hat{y}_3\hat{y}_4 - 1.019991 \times 10^{-2}\hat{y}_2^2 \\ &+ 3.928872 \times 10^{-2}\hat{y}_3^2 \\ &- 1.961657 \times 10^{-2}\hat{y}_4^2 \end{aligned}$$

For comparisons the solutions of eqs (106) (original model) and (111) (approximately lumped model) for different initial values are given in Figs 1-3. Table 2 presents the detailed numbers for \hat{y}_1 with one initial condition to provide a quantitative comparison for all chosen initial conditions. When \hat{n} is larger, the accuracy becomes better. However, even if $\hat{n} = 2$, the error is still quite small.

6B. Example 2

The second example is the same system except that $x_1 = 0.1$. In this case, the eigenvalues of A_0 are -1.1 , -2 , -2 , -2 , $-(1 + \sqrt{2})$, $-(1 + \sqrt{2})$, 0 and 0 . We cannot ignore the difference between -1.1 and -2 . Therefore, the expanded root subspace corresponding to -1.1 and -2 cannot be used in this case. The other procedures are the same as in example 1. All the exact lumping schemes in example 1 can be obtained in example 2. For the same M_0 as that of example 1 the resultant approximate lumping matrices validated in the whole composition region with different \hat{n} and the corresponding Z_1 and Z_2 are given below:

\hat{n}	2	3	4
Z_1	0.82	0.57	0.36
Z_2	0.06	0.22	0.38

$$N = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.2945 & 0.6025 & 0.5220 & 0.5222 & -0.0017 & 0.0271 & -0.0577 & 0.0324 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.3196 & 0.5953 & 0.5205 & 0.5199 & -0.0297 & 0.0259 & -0.0163 & 0.0201 \\ 0.8486 & -0.5237 & 0.0427 & 0.0304 & -0.0315 & 0.0301 & -0.0224 & 0.0238 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5389 & 0.4324 & 0.5427 & 0.4750 & -0.0334 & 0.0275 & -0.0130 & 0.0189 \\ 0.5304 & -0.4455 & 0.3710 & -0.6182 & 0.0080 & 0.0149 & 0.0101 & -0.0031 \\ 0.5537 & -0.4135 & -0.5877 & 0.4207 & 0.0029 & -0.0091 & 0.0069 & -0.0007 \end{pmatrix}$$

Lumped kinetic equations with $\hat{n} = 2$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.808657\hat{y}_2 \\ d\hat{y}_2/dt &= 1.269979 \times 10^{-2}\hat{y}_1 - 1.880187\hat{y}_2 \\ &\quad - 0.108259\hat{y}_2^2 \end{aligned} \quad (112)$$

Lumped kinetic equations with $\hat{n} = 3$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.811499\hat{y}_2 + 1.61235 \times 10^{-2}\hat{y}_3 \\ d\hat{y}_2/dt &= -3.988877 \times 10^{-3}\hat{y}_1 - 1.902552\hat{y}_2 \\ &\quad + 0.261252\hat{y}_3 + 6.570965 \times 10^{-2}\hat{y}_2\hat{y}_3 \\ &\quad - 8.808404 \times 10^{-2}\hat{y}_2^2 - 0.112177\hat{y}_2^3 \quad (113) \\ d\hat{y}_3/dt &= -3.183231 \times 10^{-3}\hat{y}_1 + 0.253534\hat{y}_2 \\ &\quad - 1.337853\hat{y}_3 - 0.131870\hat{y}_2\hat{y}_3 \\ &\quad + 0.176772\hat{y}_2^2 + 0.225123\hat{y}_2^3 \end{aligned}$$

Lumped kinetic equations with $\hat{n} = 4$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.746556\hat{y}_2 - 0.401056\hat{y}_3 - 0.275967\hat{y}_4 \\ d\hat{y}_2/dt &= -6.020594 \times 10^{-3}\hat{y}_1 - 1.729282\hat{y}_2 \\ &\quad + 0.276044\hat{y}_3 - 0.271043\hat{y}_4 + 2.854569 \\ &\quad \times 10^{-2}\hat{y}_2\hat{y}_3 + 1.096059 \times 10^{-2}\hat{y}_2\hat{y}_4 \end{aligned}$$

$$\begin{aligned} &-0.139489\hat{y}_3\hat{y}_4 - 2.618984 \times 10^{-2}\hat{y}_2^2 \\ &+ 2.061659 \times 10^{-2}\hat{y}_3^2 \\ &+ 2.461457 \times 10^{-2}\hat{y}_4^2 \\ d\hat{y}_3/dt &= 2.002723 \times 10^{-3}\hat{y}_1 + 0.251290\hat{y}_2 \\ &- 1.755447\hat{y}_3 + 0.278558\hat{y}_4 \\ &- 0.704581\hat{y}_2\hat{y}_3 - 7.855206 \times 10^{-2}\hat{y}_2\hat{y}_4 \\ &+ 0.999685\hat{y}_3\hat{y}_4 + 0.187697\hat{y}_2^2 \\ &- 0.147754\hat{y}_3^2 - 0.176407\hat{y}_4^2 \quad (114) \\ d\hat{y}_4/dt &= 8.988434 \times 10^{-4}\hat{y}_1 + 0.264633\hat{y}_2 \\ &+ 0.259676\hat{y}_3 - 1.712459\hat{y}_4 \\ &- 0.189171\hat{y}_2\hat{y}_3 - 7.263547 \times 10^{-2}\hat{y}_2\hat{y}_4 \\ &+ 0.924388\hat{y}_3\hat{y}_4 + 0.173559\hat{y}_2^2 \\ &- 0.136625\hat{y}_3^2 - 0.163120\hat{y}_4^2 \end{aligned}$$

For comparison the solutions of eqs (106) (original model) and (114) (approximately lumped model) for different initial values are given in Figs 4-6. Table 3 provides a quantitative comparison of \hat{y}_1 with one

Table 2. Comparison of solutions of \hat{y}_1 by eqs (106) and (109)-(111) [the initial concentrations are $y_1(0) = y_4(0) = 0.5$, others are zero]

t	Equation (106) (exact)	Equation (111) ($\hat{n} = 4$)	Equation (110) ($\hat{n} = 3$)	Equation (109) ($\hat{n} = 2$)
0.0	0.0000	0.0000	0.0000	0.0000
0.2	0.1615	0.1614	0.1611	0.1472
0.4	0.2708	0.2706	0.2698	0.2493
0.6	0.3447	0.3446	0.3430	0.3202
0.8	0.3948	0.3946	0.3925	0.3694
1.0	0.4288	0.4284	0.4258	0.4036
1.4	0.4673	0.4667	0.4636	0.4439
1.8	0.4850	0.4842	0.4808	0.4635
2.2	0.4931	0.4921	0.4888	0.4731
2.6	0.4968	0.4957	0.4926	0.4778
3.0	0.4985	0.4972	0.4945	0.4802
4.0	0.4998	0.4980	0.4963	0.4825
5.0	0.5000	0.4978	0.4972	0.4834

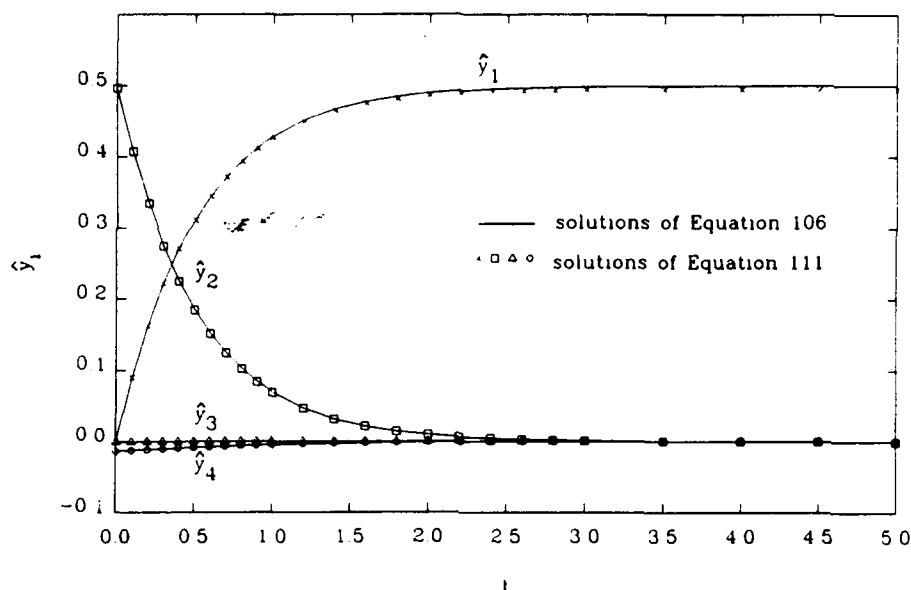


Fig. 1. Comparison between the solutions of eqs (106) and (111) [initial condition: $y_1(0) = y_2(0) = 0.5$, others are zero].

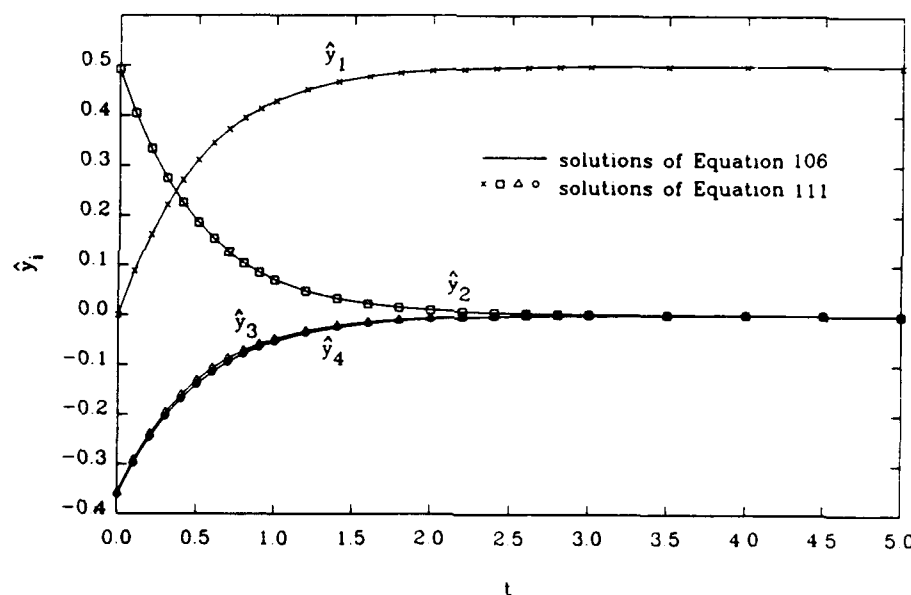


Fig. 2. Comparison between the solutions of eqs (106) and (111) [initial condition: $y_1(0) = y_4(0) = 0.5$, others are zero].

initial condition for different n . The error is larger than example 1. This is due to the larger change of k_{s1} . From Figs 4-6 one can see that the approximate lumping scheme in eq. (114) is quite good for the initial condition $y_1(0) = y_4(0) = 0.5$, but is not as good for other initial conditions. This is not surprising, because the lumping scheme is obtained in the whole composition region. If we determine the lumping scheme in a small region, the accuracy will be better.

From these examples one can see that the approach presented in this paper is capable of producing exact lumping schemes, when they exist, as well as acceptable approximate lumping ones in the presence of

constraints. This work shows that the analysis of approximate lumping is general and the suggested approach is applicable to other complicated reaction systems and other problems.

7. CONCLUSION AND DISCUSSION

In the present paper, a general analysis of approximate lumping is presented. Our previous exact lumping analysis was employed as a rigorous starting point. A general approach to construct the kinetic equations of the approximately lumped system was developed. This method can be applied to any reaction system or other kinetic systems described by a set

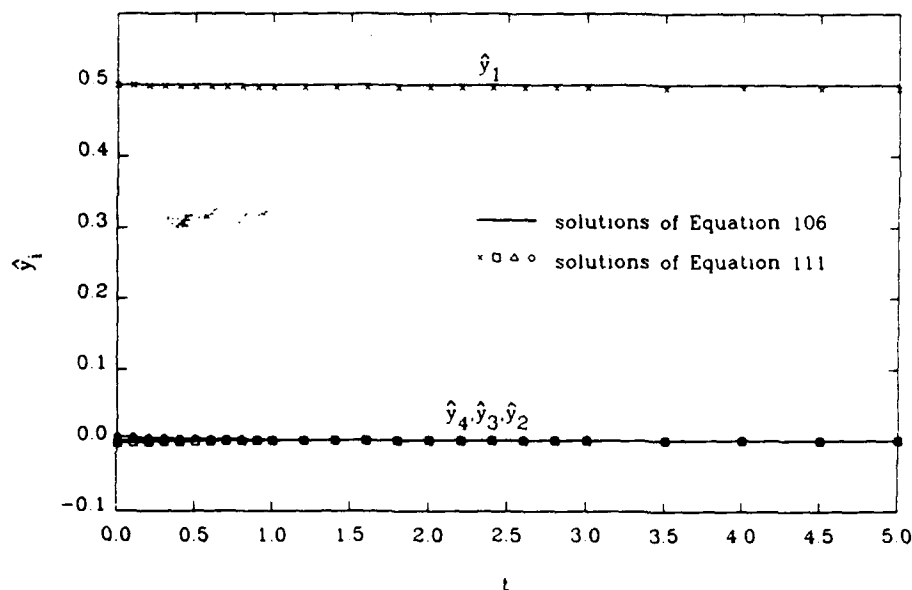


Fig. 3. Comparison between the solutions of eqs (106) and (111) [initial condition: $y_5(0) = y_7(0) = 0.5$, others are zero].

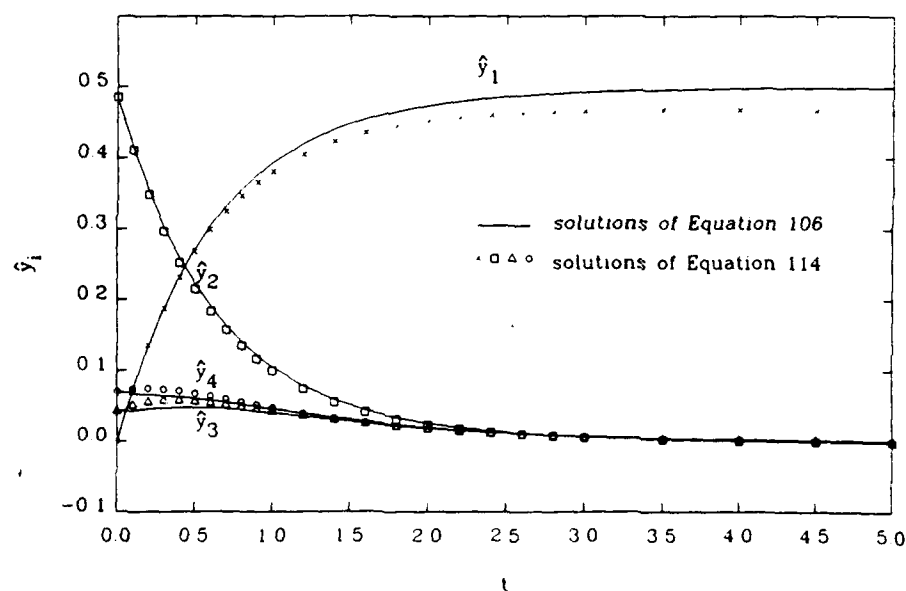


Fig. 4. Comparison between the solutions of eqs (106) and (114) [initial condition: $y_1(0) = y_2(0) = 0.5$, others are zero].

of first-order ordinary differential equations with arbitrary nonlinear coupling.

The observer theory initiated by Luenberger was formally employed to obtain the kinetic equations of the approximately lumped system. These kinetic equations have the same form as that of the exact lumped one. The difference between the approximately lumped kinetic equations and those of an exactly lumped system is that now the equations are dependent on the generalized inverse of the lumping matrix. If we are only concerned about the error and do not require the lumped system to follow uni- and/

or bimolecular reaction schemes and other restrictions, a good choice of the generalized inverse of the lumping matrix is the $\{1, 2, 3, 4\}$ -inverse. When the rows of the lumping matrix are orthonormal, it is simply M^T .

Using the results of our exact lumping analysis the equations were derived which can be applied to obtain the approximate lumping matrices with or without physical constraints. These equations can be employed to determine the approximate lumping schemes in the entire composition region or only in a small region of it, or even along a reaction path. The

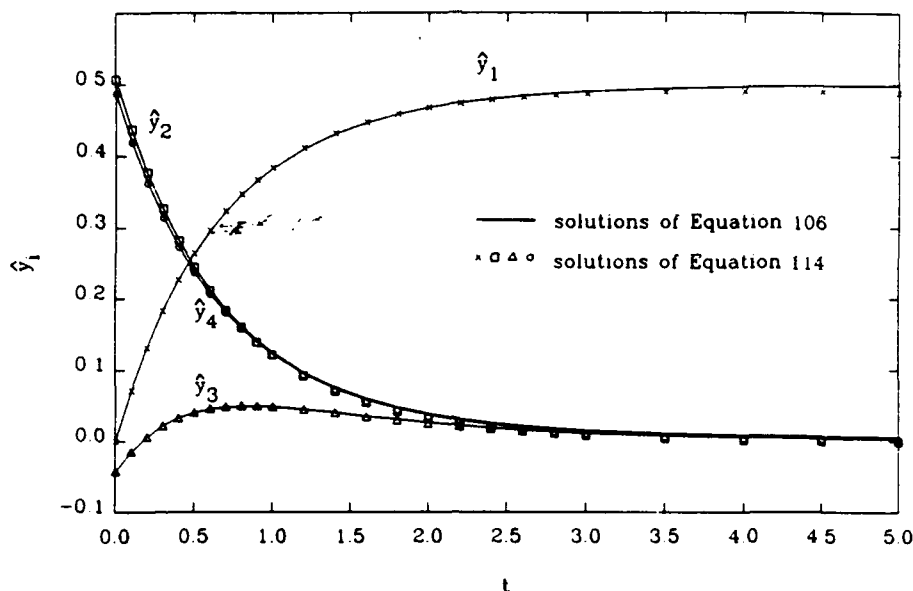


Fig. 5. Comparison between the solutions of eqs (106) and (114) [initial condition: $y_1(0) = y_4(0) = 0.5$, others are zero].

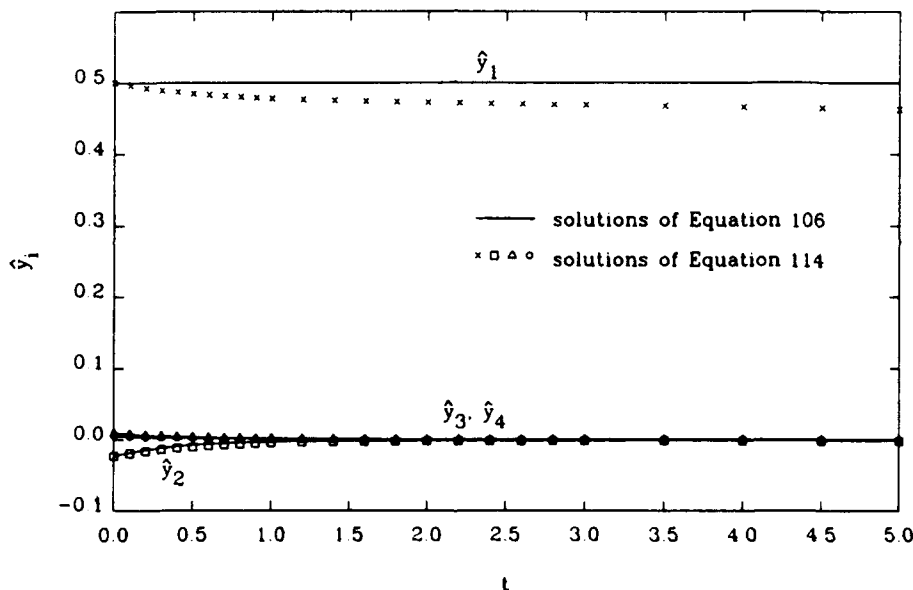


Fig. 6. Comparison between the solutions of eqs (106) and (114) [initial condition: $y_5(0) = y_7(0) = 0.5$, others are zero].

equations are invariant to the different regions of the composition region, but the parameters in the equations depend on the region; especially for a reaction path, they depend on the initial value $y(0)$. The equations to calculate these parameters were presented.

In order to reach the global minimum solutions of the equations an approach to choose suitable initial M values was developed. This approach is based on the concept of the degree of coincidence between the invariant subspaces of A_k s. A global minimum solution is located in a subspace spanned by the basis vectors of the set of A_k -invariant subspaces with the largest sum of degrees of coincidence. An example

modified from a case of exact lumping was employed to examine this method.

The approach presented here for constructing the approximately lumped kinetic equations is quite general. It is applicable to many reaction systems or other problems, such as in chemical engineering, control problems or even classical molecular mechanics. However, this method is specifically suitable for uni- and/or bimolecular reaction systems, because the transpose of the Jacobian matrix of these systems is readily decomposed into a certain linear combination of constant matrices. For other systems we need to find an easy way to do so. The same problem also

Table 3. Comparison of solutions of \hat{y}_1 by eqs (106) and (112)–(114) [the initial concentrations are $y_1(0) = y_4(0) = 0.5$, others are zero]

t	Equation (106) (exact)	Equation (114) ($\hat{n} = 4$)	Equation (113) ($\hat{n} = 3$)	Equation (112) ($\hat{n} = 2$)
0.0	0.0000	0.0000	0.0000	0.0000
0.2	0.1318	0.1315	0.1305	0.1227
0.4	0.2267	0.2268	0.2246	0.2065
0.6	0.2955	0.2963	0.2929	0.2640
0.8	0.3458	0.3470	0.3429	0.3035
1.0	0.3829	0.3843	0.3798	0.3308
1.4	0.4312	0.4323	0.4272	0.3631
1.8	0.4587	0.4590	0.4543	0.3792
2.2	0.4747	0.4739	0.4693	0.3878
2.6	0.4843	0.4822	0.4778	0.3929
3.0	0.4902	0.4867	0.4824	0.3963
4.0	0.4968	0.4899	0.4862	0.4022
5.0	0.4990	0.4887	0.4856	0.4073

appears for nonisothermal reaction systems, whose rate constants are functions of temperature. Therefore, refining the present approach to stronger nonlinearities is an important task.

When the dimension of the original system is high, the determination of the initial values of the matrix equation for M becomes very expensive by using the degree of the coincidence of the invariant subspaces of A_k s. This restricts the application of the present approach. Fortunately, we will prove in another paper that the necessary and sufficient condition for exact lumping validated in the Y_n -space is only the invariance of \mathcal{M} to $J^T(y)$ without the requirement of the equality of the eigenvalues of \mathcal{M} to $J^T(y)$ and $J^T(\bar{M}My)$, or alternatively the representation in eq. (19). This reduced requirement simplifies the determination of the approximate lumping schemes. We have accordingly developed an easy way to determine the constrained lumping schemes validated in the Y_n -space. The resultant M can also be employed as an initial value of the matrix equations to find the approximate lumping schemes in any desired region.

Acknowledgements—The authors acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

NOTATION

Scalars

$a_k(y)$	k th coefficient of the decomposition of $J^T(y)$
a_{kk}	defined as $\int_{\Omega} a_k(y) a_{k'}(y) d\Omega$
$b_{kk}(M)$	defined as $\int_{\Omega} a_k(y) a_{k'}(z) d\Omega$
$c_{kk}(M)$	defined as $\int_{\Omega} a_k(z) a_{k'}(z) d\Omega$
c	upper limit of total concentration
C_i	i th species of a reaction system
d_c	degree of coincidence of two subspaces

D_c	defined $\sum_{\substack{k, k'=1 \\ k < k'}}^m d_c(k, k')$
k	integer
l	integer
m	integer
m_{kl}	(k, l) -entry of M
\mathcal{M}	subspace
$\mathcal{M}(r)$	subspace with dimension r
n	dimension of vector y
\hat{n}	dimension of vector \hat{y}
r	integer
s	trajectory, length of the reaction path in composition space, dummy variable or an integer
s_f	final value of the length of a reaction path
S	sum of degrees of coincidence
S_G	defined as $\max_{M_G M_G^T = I} \text{tr } M_G Y M_G^T$
S_D	defined as $\max_{M_D M_D^T = I} \text{tr } M_D Y M_D^T$
S_1	n -component kinetic system
S_2	\hat{n} -component kinetic system driven by S_1
t	time
y_k	k th element of vector y
Y_n	n -dimensional composition space
$Y_{\hat{n}}$	\hat{n} -dimensional composition subspace
Z_1	total error defined as $\text{tr} \sum_{k, k'=1}^m a_{kk} M A_k^T (I_n - M^T M) A_k M^T$
Z_2	total error defined as $\text{tr} \sum_{k, k'=1}^m \beta_{kk}(M) M A_k^T A_k M^T$
Z'_1	objective function
$Z_1(y)$	defined as $\text{tr} [E_1^T(y) E_1(y)]$
$Z_2(y)$	defined as $\text{tr} [E_2^T(y) E_2(y)]$
$Z(\bar{M}, y)$	defined as $\text{tr} [E_1^T(\bar{M}, y) E_1(\bar{M}, y)]$

Vectors and matrices

Capital letters represent matrices, bold-face lower-case letters represent vectors.

A	constant matrix
A_0	constant matrix
A_k	constant matrix
B	constant matrix
e_i	unit vector with 1 as its i th element, and 0 for the rest of the elements
$e(y)$	error vector
$E_1(y)$	error matrix defined as $(I_n - M^T \bar{M}^T) J^T(y) M^T$
$E_1(\bar{M}, y)$	error matrix defined as $(I_n - M^T \bar{M}^T) J^T(y) M^T$ with a given M
$E_2(y)$	error matrix defined as $M[J(y) - J(\bar{M}My)]$
$f(y)$	n -dimensional function vector
$\hat{f}(\hat{y})$	\hat{n} -dimensional function vector
$F(X)$	function matrix
$G(X)$	function matrix
H	permutation matrix
I	identity matrix
$J(y)$	Jacobian matrix of $f(y)$
m_i	column i of M
M	lumping matrix
M_D	determined submatrix of M
M_G	given submatrix of M
\bar{M}	generalized inverse of M satisfying $M\bar{M} = I_n$
M^+	$\{1, 2, 3, 4\}$ -generalized inverse of M
P	constant matrix
P_k	constant matrix
Q	$\hat{n} \times \hat{n}$ matrix
$Q(y)$	$\hat{n} \times \hat{n}$ function matrix
R	eigenvector matrix of Y
$R(i)$	eigenvector matrix of $Y(i)$
x	n -dimensional vector
X_{A_k}	eigenvector matrix of A_k
y	n -dimensional variable vector
\bar{y}	defined as $\bar{M}_i My$
\hat{y}	\hat{n} -dimensional variable vector
Y	defined as $\sum_{k=1}^m \frac{1}{\hat{n}} Y_k(r_k) Y_k(r_k)^T$
$Y(i)$	matrix Y for the i th closest group
$Y(r)$	$n \times r$ matrix with orthonormal columns
$\bar{Y}(r)$	$n \times r$ matrix with orthonormal columns
$Y_k(i)$	i th submatrix of X_{A_k}
W	$n \times (n - r)$ matrix, which is orthogonal to $Y(r)$
Z	$n \times \hat{n}$ arbitrary matrix
z	defined as $\bar{M}My$

Greek letters

α_i	constant vector
$\beta_{kk}(M)$	defined as $a_{kk} - b_{kk}(M) - b_{k^*k}(M) + c_{kk}(M)$
δ_{ij}	Kronecker delta function with value 1 for $i = j$, 0 for $i \neq j$
λ_i	i th eigenvalue of matrix A_k or Y
Λ	Lagrange multiplier matrix with λ_{ij} as its (i, j) -entry
Ω	desired region of Y_n -space

Ω_i	defined as $\bar{M}_i M \Omega$
Ω_{total}	defined as $\cup_{i=0}^{\infty} \bar{M}_i M \Omega$

Symbols

\sim	any property related to the lumped system
$*$	any property related to stable state
\otimes	Kronecker product of matrices
0	null vector
0	null matrix

REFERENCES

- Bellman, R., 1970, *Introduction to Matrix Analysis*. McGraw-Hill, New York.
- Ben-Israel, A. and Greville, T. N. E., 1974, *Generalized Inverse: Theory and Applications*. John Wiley, New York.
- Golikeri, S. V. and Luss, D., 1972, Analysis of activation energy of grouped parallel reactions. *A.I.Ch.E. J.* **18**, 277-282.
- Golikeri, S. V. and Luss, D., 1974, Aggregation of many coupled consecutive first order reactions. *Chem. Engng Sci.* **29**, 845-855.
- Hutchinson, P. and Luss, D., 1970, Lumping of mixtures with many parallel first order reactions. *Chem. Engng J.* **1**, 129-135.
- Kuo, J. C. W. and Wei, J., 1969, A lumping analysis in monomolecular reaction systems—analysis of approximately lumpable system. *Ind. Engng Chem. Fundam.* **8**, 124-133.
- Li, G. and Rabitz, H., 1989, A general analysis of exact lumping in chemical kinetics. *Chem. Engng Sci.* **44**, 1413-1430.
- Liu, Y. A. and Lapidus, L., 1973, Observer theory for lumping analysis of monomolecular reaction systems. *A.I.Ch.E. J.* **19**, 467-473.
- Luenberger, D. G., 1964, Observing the state of a linear system. *IEEE Trans. Mil. Electron.* **MIL-8***, 74-80.
- Luss, D. and Hutchinson, P., 1971, Lumping of mixture with many parallel N -th order reactions. *Chem. Engng J.* **2**, 172-177.
- Luss, D., 1972, Grouping of many species each consumed by two parallel first-order reactions. *A.I.Ch.E. J.* **21**, 865-872.
- Wei, J. and Kuo, J. C. W., 1969, A lumping analysis in monomolecular reaction systems—analysis of exactly lumpable system. *Ind. Engng Chem. Fundam.* **8**, 114-123.
- Yetter, R. A., Dryer, F. F. and Rabitz, H., 1985, Some interpretive aspects of elementary sensitivity gradients in combustion kinetics modelling. *Combust. Flame* **59**, 107-133.

APPENDIX A: DERIVATION OF EQ. (54)

In order to determine the M which gives the smallest error, we need to minimize the function Z'_1 with respect to M^T and λ_{ij} (for all i and j). This means that we need to solve the following equations:

$$\begin{aligned} \partial Z'_1 / \partial M^T &= 0, \\ \partial Z'_1 / \partial \lambda_{ij} &= 0 \quad (\text{for all } i \text{ and } j) \end{aligned} \quad (\text{A1})$$

where the function Z'_1 is

$$\begin{aligned} Z'_1 &= \text{tr} \sum_{k, k^*=1}^m a_{kk} M A_k^T (I_n - M^T M) A_k M^T \\ &+ \sum_{i, j=1}^{\hat{n}} \lambda_{ij} \left(\sum_{s=1}^n m_{is} m_{js} - \delta_{ij} \right). \end{aligned} \quad (\text{A2})$$

and λ_{ij} are Lagrange multipliers. The first equation of eq.

(A1) can be written as follows:

$$\begin{aligned} \partial Z_1' / \partial M^T = & \sum_{k,k'=1}^n a_{kk'} \left[\frac{\partial}{\partial M^T} \text{tr}(MA_k^T A_{k'} M^T) \right. \\ & \left. - \frac{\partial}{\partial M^T} \text{tr}(MA_k^T M^T MA_{k'} M^T) \right] \\ & + \frac{\partial}{\partial M^T} \sum_{i,j=1}^n \lambda_{ij} \left(\sum_{s=1}^n m_{is} m_{js} - \delta_{ij} \right). \end{aligned} \quad (\text{A3})$$

Since this equation involves the derivatives of a matrix with respect to a matrix, we state some relevant results (see any textbook on matrix calculus), which will be used in the following calculations.

$$(1) \partial F[G(X)] / \partial X = [\partial G(X) / \partial X] [\partial F(G) / \partial G] \quad (\text{A4})$$

$$(2) \partial [F(X)G(X)] / \partial X = [\partial F(X) / \partial X] (I_p \otimes G(X)) + [\partial G(X) / \partial X] [F^T(X) \otimes I_r] \quad (\text{A5})$$

$$(3) \partial \text{tr}(AX) / \partial X = A^T \quad (\text{A6})$$

$$(4) \partial \text{tr}(X^T AX) / \partial X = (A + A^T)X \quad (\text{A7})$$

$$(5) \text{vec } A = (a_{11} \ a_{12} \ \dots \ a_{1n} \ a_{21} \ \dots \ a_{mn})^T \quad (\text{A8})$$

$$(6) \text{vec}(AXB) = (A \otimes B^T) \text{vec } X \quad (\text{A9})$$

The symbol \otimes denotes a Kronecker product and $F(X)$, $G(X)$ are $p \times q$ and $q \times r$ matrices, respectively.

We will determine each term of eq. (A3) separately. Let

$$A = A_k^T A_{k'}. \quad (\text{A10})$$

Using eq. (A7) we obtain

$$\begin{aligned} \partial \text{tr}(MA_k^T A_{k'} M^T) / \partial M^T &= (A + A^T)M^T \\ &= (A_k^T A_{k'} + A_{k'}^T A_k)M^T. \end{aligned} \quad (\text{A11})$$

Let

$$Z_{kk'} = MA_k^T M^T MA_{k'} M^T. \quad (\text{A12})$$

Then

$$\begin{aligned} \partial \text{tr } Z_{kk'} / \partial M^T &= \frac{\partial Z_{kk'}}{\partial M^T} \frac{\partial \text{tr } Z_{kk'}}{\partial Z_{kk'}} \\ &= \frac{\partial Z_{kk'}}{\partial M^T} \text{vec } I_n. \end{aligned} \quad (\text{A13})$$

$$\begin{aligned} \partial Z_{kk'} / \partial M^T &= \partial (MA_k^T M^T MA_{k'} M^T) / \partial M^T \\ &= \partial [F(M^T)G(M^T)] / \partial M^T \\ &= \frac{\partial F(M^T)}{\partial M^T} [I_n \otimes G(M^T)] \\ &\quad + \frac{\partial G(M^T)}{\partial M^T} [F^T(M^T) \otimes I_n] \end{aligned} \quad (\text{A14})$$

where

$$F(M^T) = MA_k^T M^T \quad (\text{A15})$$

$$G(M^T) = MA_{k'} M^T. \quad (\text{A16})$$

Utilizing the appropriate equations of matrix calculus we can determine all the terms in eq. (A14):

$$\begin{aligned} \partial F(M^T) / \partial M^T &= \partial (MA_k^T M^T) / \partial M^T \\ &= \frac{\partial M}{\partial M^T} (I_n \otimes A_k^T M^T) + \frac{\partial A_k^T M^T}{\partial M^T} (M^T \otimes I_n) \\ &= H(I_n \otimes A_k^T M^T) + (A_k \otimes I_n)(M^T \otimes I_n) \\ &= H(I_n \otimes A_k^T M^T) + (A_k M^T \otimes I_n) \end{aligned} \quad (\text{A17})$$

where H is known as a permutation matrix satisfying

$$H \text{vec } M^T = \text{vec } M \quad (\text{A18})$$

and

$$\begin{aligned} \partial G(M^T) / \partial M^T &= \partial (MA_{k'} M^T) / \partial M^T \\ &= \frac{\partial M}{\partial M^T} (I_n \otimes A_{k'} M^T) + \frac{\partial A_{k'} M^T}{\partial M^T} (M^T \otimes I_n) \\ &= H(I_n \otimes A_{k'} M^T) + (A_{k'}^T \otimes I_n)(M^T \otimes I_n) \\ &= H(I_n \otimes A_{k'} M^T) + (A_{k'}^T M^T \otimes I_n). \end{aligned} \quad (\text{A19})$$

Substituting eqs (A17) and (A19) into eq. (A14) we obtain

$$\begin{aligned} \partial Z_{kk'} / \partial M^T &= [H(I_n \otimes A_k^T M^T) + (A_k M^T \otimes I_n)] \\ &\quad \times (I_n \otimes MA_{k'} M^T) \\ &\quad + [H(I_n \otimes A_{k'} M^T) + (A_{k'}^T M^T \otimes I_n)] \\ &\quad \times (MA_k M^T \otimes I_n) \\ &= H(I_n \otimes A_k^T M^T MA_{k'} M^T) \\ &\quad + (A_k M^T \otimes MA_{k'} M^T) \\ &\quad + H(MA_k M^T \otimes A_{k'} M^T) \\ &\quad + (A_k^T M^T MA_{k'} M^T \otimes I_n). \end{aligned} \quad (\text{A20})$$

Then substituting eq. (A20) into eq. (A13) gives

$$\begin{aligned} \partial \text{tr } Z_{kk'} / \partial M^T &= H(I_n \otimes A_k^T M^T MA_{k'} M^T) \text{vec } I_n \\ &\quad + (A_k M^T \otimes MA_{k'} M^T) \text{vec } I_n \\ &\quad + H(MA_k M^T \otimes A_{k'} M^T) \text{vec } I_n \\ &\quad + (A_k^T M^T MA_{k'} M^T \otimes I_n) \text{vec } I_n \\ &= H \text{vec } (A_k^T M^T MA_{k'} M^T)^T \\ &\quad + \text{vec } (A_k M^T MA_{k'}^T M^T) \\ &\quad + H \text{vec } (MA_k M^T MA_{k'}^T) \\ &\quad + \text{vec } (A_k^T M^T MA_{k'} M^T). \end{aligned}$$

Representing this result in the form of matrix we obtain

$$\begin{aligned} \partial \text{tr } Z_{kk'} / \partial M^T &= A_k^T M^T MA_{k'} M^T + A_k M^T MA_{k'}^T M^T \\ &\quad + A_{k'} M^T MA_k^T M^T + A_{k'}^T M^T MA_{k'} M^T. \end{aligned} \quad (\text{A21})$$

For the last term in eq. (A3) we first consider differentiation with respect to the element m_{kl} of M :

$$\begin{aligned} \frac{\partial}{\partial m_{kl}} \sum_{i,j=1}^n \lambda_{ij} \left(\sum_{s=1}^n m_{is} m_{js} - \delta_{ij} \right) &= 2 \sum_{i=1}^n \lambda_{ik} m_{il} \\ &= 2m_l \lambda_k \end{aligned} \quad (\text{A22})$$

where m_l represents the l th column of M and $\lambda_k = (\lambda_{1k}, \lambda_{2k}, \dots, \lambda_{nk})^T$. Here we used the property of $\lambda_{ij} = \lambda_{ji}$. Then the differentiation of the last term with respect to M^T can be described as follows:

$$\frac{\partial}{\partial M^T} \sum_{i,j=1}^n \lambda_{ij} \left(\sum_{s=1}^n m_{is} m_{js} - \delta_{ij} \right) = 2M^T \Lambda \quad (\text{A23})$$

where Λ is the matrix whose (i, j) -entry is λ_{ij} .

Substituting all the results into eq. (A3) gives

$$\begin{aligned} \partial Z_1' / \partial M^T &= \sum_{k,k'=1}^n a_{kk'} (A_k^T A_{k'} + A_{k'}^T A_k - A_k^T M^T MA_{k'} \\ &\quad - A_{k'} M^T MA_k^T - A_{k'}^T M^T MA_k M^T + 2M^T \Lambda) \\ &= 2 \sum_{k,k'=1}^n a_{kk'} (A_k^T A_{k'} - A_k^T M^T MA_{k'} \\ &\quad - A_{k'} M^T MA_k^T) M^T + 2M^T \Lambda = 0. \end{aligned}$$

This gives the following equation:

$$\sum_{k,k'=1}^m a_{kk'} (A_k^T A_{k'} - A_k^T M^T M A_{k'}) - A_k M^T M A_{k'}^T M^T + M^T \Lambda = 0. \quad (\text{A24})$$

Now consider the differentiation of Z'_1 with respect to Λ . It is easy to show that the result is

$$\partial Z'_1 / \partial \Lambda = M M^T - I_n = 0.$$

This gives the restriction condition:

$$M M^T = I_n. \quad (\text{A25})$$

Multiplying both sides of eq. (A24) from the left by M we obtain

$$\sum_{k,k'=1}^m a_{kk'} M (A_k^T A_{k'} - A_k^T M^T M A_{k'}) - A_k M^T M A_{k'}^T M^T + \Lambda = 0.$$

Therefore Λ can be expressed as

$$\Lambda = - \sum_{k,k'=1}^m a_{kk'} M (A_k^T A_{k'} - A_k^T M^T M A_{k'}) - A_k M^T M A_{k'}^T M^T. \quad (\text{A26})$$

Substituting it into eq. (A24) we obtain the final result:

$$(I_n - M^T M) \sum_{k,k'=1}^m a_{kk'} (A_k^T A_{k'} - A_k^T M^T M A_{k'}) - A_k M^T M A_{k'}^T M^T = 0, \quad (\text{A27})$$

which is eq. (54) in the text.

APPENDIX B: THE DEGREE OF COINCIDENCE BETWEEN TWO SUBSPACES

We need to give a quantitative description of the degree of coincidence between two subspaces. We use d_c to represent it. According to the geometric concept, when one of the two subspaces is inside the other one, d_c is unity; when the two subspaces are orthogonal to each other, $d_c = 0$. In other cases, $0 < d_c < 1$. d_c should also be independent of the bases for the two subspaces.

Suppose $\mathcal{M}(r)$ and $\mathcal{M}(r')$ are r - and r' -dimensional subspaces, respectively. We choose corresponding r and r' orthonormal vectors as their bases. Let the $n \times r$ and $n \times r'$ matrices $Y(r)$ and $Y(r')$ be the matrix representations of the two subspaces with $r' \leq r$. If the degree of coincidence d_c of the two subspaces is defined as follows:

$$d_c = \frac{1}{r'} \text{tr} [Y(r')^T Y(r) Y(r)^T Y(r')] \quad (\text{B1})$$

we can prove that d_c satisfies the above requirements.

First, when one of the two subspaces is inside the other one, i.e. the basis vectors of a subspace are certain linear combinations of those in the other subspace, we can prove that d_c is equal to unity. In this case the columns of $Y(r')$ are linear combinations of those of $Y(r)$, and then we have

$$Y(r')_i = Y(r) \alpha_i \quad (\text{B2})$$

where $Y(r')_i$ is the i th column of $Y(r')$ and α_i is a r -dimensional vector. Since $Y(r')_i$ is normalized, then

$$\begin{aligned} Y(r')_i^T Y(r')_i &= \alpha_i^T Y(r)^T Y(r) \alpha_i \\ &= \alpha_i^T \alpha_i = 1. \end{aligned} \quad (\text{B3})$$

This shows that α_i is also a normalized vector. According to

eq. (B1) in this case we have

$$\begin{aligned} d_c &= \frac{1}{r'} \text{tr} [Y(r')^T Y(r) Y(r)^T Y(r')] \\ &= \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r [Y(r')_i^T Y(r)_j]^2 \\ &= \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r [\alpha_i^T Y(r)^T Y(r)_j]^2 \\ &= \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r (\alpha_i^T e_j)^2 \\ &= \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r (\alpha_{ij})^2 \\ &= \frac{1}{r'} \sum_{i=1}^{r'} 1 \\ &= \frac{1}{r'} r' = 1 \end{aligned} \quad (\text{B4})$$

where e_j is a unit vector with its j th element 1, the rest 0, and α_{ij} is the j th element of α_i .

As another case consider the two subspaces as being orthogonal to each other. In this case, we have

$$Y(r')^T Y(r) = 0. \quad (\text{B5})$$

The degree of coincidence between these two subspaces is

$$d_c = \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r [Y(r')_i^T Y(r)_j]^2 = \frac{1}{r'} \sum_{i=1}^{r'} \sum_{j=1}^r 0 = 0. \quad (\text{B6})$$

In general we can prove that the degree of coincidence of two arbitrary subspaces is between 0 and 1. Notice that the sum of the degrees of coincidence between vector $Y(r')_i$ and all columns of $Y(r)$ can be obtained as follows:

$$Y(r')_i^T Y(r) Y(r)^T Y(r')_i = Y(r')_i^T (I_n - W W^T) Y(r')_i \quad (\text{B7})$$

where W is an $n \times (n - r)$ matrix, which is orthogonal to $Y(r)$ and its columns are orthonormal. We know that for an $n \times n$ symmetric matrix A

$$\max_{|x|=1} x^T A x = \lambda_1(A) \quad (\text{B8})$$

$$\min_{|x|=1} x^T A x = \lambda_n(A) \quad (\text{B9})$$

where $\lambda_1(A)$ and $\lambda_n(A)$ are the largest and the smallest eigenvalues of A (Bellman, 1970). We also know that for a nonnegative definite matrix the eigenvalues are nonnegative. $Y(r) Y(r)^T$ is nonnegative definite, so its eigenvalues are equal to or larger than zero. This means that

$$Y(r')_i^T Y(r) Y(r)^T Y(r')_i \geq 0.$$

$$d_c = \frac{1}{r'} \sum_{i=1}^{r'} Y(r')_i^T Y(r) Y(r)^T Y(r')_i \geq 0. \quad (\text{B10})$$

Considering that $Y(r) Y(r)^T$, I_n and $W W^T$ are all nonnegative definite, I_n and $W W^T$ can be diagonalized simultaneously, and

$$Y(r) Y(r)^T = I_n - W W^T \quad (\text{B11})$$

so we have

$$\lambda_i [Y(r) Y(r)^T] = \lambda_i(I_n) - \lambda_i(W W^T). \quad (\text{B12})$$

Then the eigenvalues of $Y(r) Y(r)^T$ must be equal to or less than the eigenvalues of I_n , which are equal to 1. Thus

$$d_c = \frac{1}{r'} \sum_{i=1}^{r'} Y(r')_i^T Y(r) Y(r)^T Y(r')_i \leq \frac{r'}{r'} = 1. \quad (\text{B13})$$

We can also prove that the resultant degree of coincidence is independent of the choice of the basis vectors if these

vectors are orthonormal. Suppose $\tilde{Y}(r)$ is another choice of $Y(r)$, then we have

$$\tilde{Y}(r) = Y(r)P \quad (\text{B14})$$

where P is a $r \times r$ constant matrix. Considering $\tilde{Y}(r)$ as also being orthonormal, it follows that

$$\begin{aligned} \tilde{Y}(r)^T \tilde{Y}(r) &= P^T Y(r)^T Y(r) P \\ &= P^T P = I_r. \end{aligned} \quad (\text{B15})$$

This implies that P is an orthogonal matrix. Then we have

$$\begin{aligned} \tilde{Y}(r) \tilde{Y}(r)^T &= Y(r) P P^T Y(r)^T \\ &= Y(r) Y(r)^T. \end{aligned} \quad (\text{B16})$$

Similarly, if $\tilde{Y}(r')$ is another choice of $Y(r')$, we also have

$$\tilde{Y}(r') \tilde{Y}(r')^T = Y(r') Y(r')^T. \quad (\text{B17})$$

Then the degree of coincidence for the new choices of the orthonormal bases is

$$\begin{aligned} d_c &= \frac{1}{r} \text{tr} [\tilde{Y}(r')^T \tilde{Y}(r) \tilde{Y}(r)^T \tilde{Y}(r')] \\ &= \frac{1}{r} \text{tr} [\tilde{Y}(r) \tilde{Y}(r)^T \tilde{Y}(r') \tilde{Y}(r')^T] \\ &= \frac{1}{r} \text{tr} [Y(r) Y(r)^T Y(r') Y(r')^T] \\ &= \frac{1}{r} \text{tr} [Y(r')^T Y(r) Y(r)^T Y(r')]. \end{aligned} \quad (\text{B18})$$

This result shows that the degree of coincidence is independent of the choice of orthonormal basis vectors. Therefore we can choose them arbitrarily.

Appendix G

7. New Approaches to Determination of Constrained Lumping Schemes for a Reaction System in the Whole Composition Space, G. Li and H. Rabitz, Chem. Eng. Sci., 45, (1990).

NEW APPROACHES TO DETERMINATION OF CONSTRAINED LUMPING SCHEMES FOR A REACTION SYSTEM IN THE WHOLE COMPOSITION SPACE

GENYUAN LI and HERSCHEL RABITZ*

Department of Chemistry, Princeton University, Princeton, NJ 08544-1009, U.S.A.

(First received 3 August 1989; accepted in revised form 12 December 1989)

Abstract—Two new approaches to the determination of constrained lumping schemes are presented. They are based on the property that the lumping schemes validated in the whole composition Y_n -space of y are only determined by the invariance of the subspace spanned by the row vectors of lumping matrix M with respect to the transpose of the Jacobian matrix $J^T(y)$ for the kinetic equations. It is proved that, when a part of a lumping matrix M_G is given, each row of the part of the lumping matrix to be determined, M_D , is certain linear combinations of a set of eigenvectors of a special symmetric matrix. This symmetric matrix is related to M_G^T and $A_k M_G^T$, where A_k are the basis matrices of $J^T(y)$. It is shown that the approximate lumping matrices containing M_G with different row number \hat{n} ($\hat{n} < n$) and global minimum errors can be determined by an optimization method. Using the concept of the minimal invariant subspace of a constant matrix over a given subspace one can directly obtain the lumping matrices containing M_G with different \hat{n} . The accuracy of these lumping matrices are shown to be satisfactory in sample calculations.

1. INTRODUCTION

Recently a bunch of papers on lumping have been published (Ho and Aris, 1987; Coxson and Bischoff, 1987a, b; Astarita and Ocone, 1988; Chou and Ho, 1988, 1989; Astarita, 1989; Aris, 1989). These works deal with both the discrete and continuous reaction systems. Our previous papers (Li and Rabitz, 1989, 1990) presented approaches to exact and approximate lumping for a reaction system in a desired region Ω of the composition Y_n -space. The original reaction system with n -components can be described by

$$dy/dt = f(y) \quad (1)$$

where y is an n -composition vector; $f(y)$ is an arbitrary n -function vector, which does not contain t explicitly. If the system can be exactly lumped by an $\hat{n} \times n$ real constant matrix M with rank \hat{n} ($\hat{n} \leq n$), then for

$$\hat{y} = My \quad (2)$$

we can find an \hat{n} -function vector $\hat{f}(\hat{y})$ such that

$$d\hat{y}/dt = \hat{f}(\hat{y}). \quad (3)$$

In the previous work a necessary and sufficient condition for the existence of exact lumping was established as the following. A reaction system is exactly lumpable if and only if the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$ has nontrivial fixed invariant subspaces \mathcal{M} and the corresponding eigenvalues of \mathcal{M} for $J^T(y)$ and $J^T(\bar{M}My)$ are equal for all y in the desired region Ω , where M^T is one of the matrix representations of \mathcal{M} and \bar{M} is one of the generalized inverses of M (Ben-Israel and Greville, 1974) satisfying

$$M\bar{M} = I_{\hat{n}}. \quad (4)$$

The exactly lumped system can be described as

$$d\hat{y}/dt = Mf(\bar{M}\hat{y}). \quad (5)$$

Here we will demonstrate that, when the lumping scheme is valid in the whole composition Y_n -space, this necessary and sufficient condition can be simplified as follows. A reaction system is exactly lumpable in the whole composition Y_n -space if and only if the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$ has nontrivial fixed invariant subspaces \mathcal{M} . This result will greatly simplify the determination of exact and approximate lumping schemes because the examination of the equality of the eigenvalues of \mathcal{M} for $J^T(y)$ and $J^T(\bar{M}My)$ is quite complicated.

2. THE CONDITION UNDER WHICH A REACTION SYSTEM IS EXACTLY LUMPABLE IN THE WHOLE COMPOSITION Y_n -SPACE

In our previous papers we have proved that the invariance of \mathcal{M} to $J^T(y)$ is a necessary condition for the existence of exact lumping in any region Ω . Now we will prove that this condition is also sufficient provided that Ω is the whole composition Y_n -space.

Suppose the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$ has a nontrivial fixed \hat{n} -dimensional invariant subspace \mathcal{M} with the $(n \times \hat{n})$ -matrix representation M^T for all y in the Y_n -space. Let the orthogonal direct complement of \mathcal{M} be \mathcal{N} in Y_n with the $[n \times (n - \hat{n})]$ -matrix representation being X . In order to simplify the discussion we choose two sets of orthonormal bases for \mathcal{M} and \mathcal{N} , i.e.

$$MM^T = I_{\hat{n}} \quad (6)$$

$$X^T X = I_{n-\hat{n}} \quad (7)$$

$$MX = 0. \quad (8)$$

*Author to whom correspondence should be addressed.

Therefore, the matrix $(X|M^T)$ is an orthogonal one and its inverse is just the transpose of itself: $\begin{pmatrix} X^T \\ M \end{pmatrix}$.

Then we have

$$\begin{pmatrix} X^T \\ M \end{pmatrix} (X|M^T) = \begin{pmatrix} X^T \\ M \end{pmatrix} \begin{pmatrix} X^T \\ M \end{pmatrix} = I_n. \quad (9)$$

For the following nonsingular linear transformation

$$z = \begin{pmatrix} X^T \\ M \end{pmatrix} y \quad (10)$$

we have the inverse transformation

$$y = (X|M^T)z \quad (11)$$

and

$$\begin{aligned} dz/dt &= \begin{pmatrix} X^T \\ M \end{pmatrix} dy/dt \\ &= \begin{pmatrix} X^T \\ M \end{pmatrix} f(y) \\ &= \begin{pmatrix} X^T \\ M \end{pmatrix} f[(X|M^T)z] \\ &= g(z). \end{aligned} \quad (12)$$

The corresponding Jacobian matrix of $g(z)$ is

$$\begin{aligned} J(z) &= \partial \left\{ \begin{pmatrix} X^T \\ M \end{pmatrix} f[(X|M^T)z] \right\} / \partial z \\ &= \begin{pmatrix} X^T \\ M \end{pmatrix} \frac{\partial}{\partial y} f(y) \frac{\partial y}{\partial z} \\ &= \begin{pmatrix} X^T \\ M \end{pmatrix} J(y) (X|M^T) \\ &= \begin{bmatrix} X^T J(y) X & X^T J(y) M^T \\ M J(y) X & M J(y) M^T \end{bmatrix}. \end{aligned} \quad (13)$$

When the subspace \mathcal{M} spanned by the row vectors of M is a fixed invariant one of $J^T(y)$ for all values of y in Y_n , i.e. a left fixed invariant subspace of $J(y)$, we have

$$M J(y) = Q(y) M \quad (14)$$

where $Q(y)$ is an $(\hat{n} \times \hat{n})$ -matrix and

$$M J(y) X = Q(y) M X = 0. \quad (15)$$

Then eq. (13) becomes

$$J(z) = \begin{bmatrix} X^T J(y) X & X^T J(y) M^T \\ 0 & M J(y) M^T \end{bmatrix}. \quad (16)$$

Since the transformation in eq. (10) is nonsingular and applicable for all values of $y \in Y_n$, this implies that its image is valid for all values of $z \in Z_n$. Thus from eq. (16) we have

$$\begin{aligned} \partial g_i(z) / \partial z_j &= 0 \\ (i &= n - \hat{n} + 1, n - \hat{n} + 2, \dots, n; \\ j &= 1, 2, \dots, n - \hat{n}) \quad \forall z \in Z_n. \end{aligned} \quad (17)$$

Equation (17) shows that $g_i(z)$ ($i = n - \hat{n} + 1, n - \hat{n} + 2, \dots, n$) do not contain the first $n - \hat{n}$ variables z_j

($j = 1, 2, \dots, n - \hat{n}$). Hence, the last \hat{n} equations in eq. (12) compose an exactly lumped model.

Now we will demonstrate that this lumped model can be represented as

$$d\hat{y}/dt = M f(\bar{M}\hat{y}). \quad (18)$$

Let

$$\hat{y} = M y. \quad (19)$$

From eq. (12) one has

$$d\hat{y}/dt = M f[(X|M^T)z]. \quad (20)$$

Taking into account that these equations do not contain z_j ($j = 1, 2, \dots, n - \hat{n}$) and considering eq. (10), eq. (20) is equivalent to

$$\begin{aligned} d\hat{y}/dt &= M f[(0|M^T)z] \\ &= M f(M^T \hat{y}). \end{aligned} \quad (21)$$

Multiplying eq. (1) from the left by M and comparing the resultant equations with eq. (21) yields

$$\begin{aligned} M f(y) &= M f(M^T \hat{y}) \\ &= M f(M^T M y). \end{aligned} \quad (22)$$

This holds for any value of $y \in Y_n$. Therefore, we can take

$$y = \bar{M} \hat{y} \quad (23)$$

then

$$\begin{aligned} M f(\bar{M} \hat{y}) &= M f(M^T M \bar{M} \hat{y}) \\ &= M f(M^T \hat{y}). \end{aligned} \quad (24)$$

Substituting eq. (24) into eq. (21) gives eq. (18).

In summary, we have proved that a system is exactly lumpable in the whole Y_n -space if and only if the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$ has nontrivial fixed invariant subspaces \mathcal{M} for all $y \in Y_n$. The lumping matrix is one of the transposes of the matrix representations of \mathcal{M} . The important issue is that the lumping scheme is valid in the whole Y_n -space. Otherwise, the conclusion would be invalid. In the previous paper (Li and Rabitz, 1989) on exact lumping example 2 of a uni- and bimolecular reaction system is a demonstration of this result. In that example we did not give any restriction on the value of y , i.e. Ω is the full Y_n -space. The eigenvalues of $J^T(y)$ and $J^T(\bar{M} M y)$ for any one of the resultant 23 types of the fixed $J^T(y)$ -invariant subspaces are equal.

3. THE DETERMINATION OF CONSTRAINED APPROXIMATE LUMPING MATRICES IN THE WHOLE COMPOSITION Y_n -SPACE

An approach to the determination of constrained approximate lumping matrices has been presented (Li and Rabitz, 1990). That approach minimizes the two errors corresponding to the invariance of \mathcal{M} to $J^T(y)$ and the equality of the corresponding eigenvalues of \mathcal{M} for $J(y)$ and $J^T(\bar{M} M y)$ in Ω . Two problems arise in the determination of the approximate lumping matrices: (1) it is not easy to minimize the second error

for the equality of the eigenvalues simultaneously with the first error, and (2) for large n and \hat{n} the determination of the initial values for iteration of the matrix equations determining M is a time-consuming task. When the lumping matrix is valid in the whole Y_n -space, we only need to consider the first error for the invariance of \mathcal{M} . Taking advantage of this situation we now develop a new optimization approach to determine the constrained lumping matrices without solving the matrix equations. It will be shown in Section 3A that the new approach is much better and easier than the original one to obtain the solution of M having the global minimum error. However, in numerical calculations, especially for large n and \hat{n} , it can be a very difficult task to reach the global minimum. On the other hand, given the approximate nature of the lumping goal, some error is acceptable. Therefore, in Section 3B we develop a direct approach, which can determine the constrained approximate lumping schemes with satisfactory accuracy. This direct approach is built on the concept of the minimal A_k -invariant subspace \mathcal{M} over a given subspace \mathcal{M}_G . This approach will directly supply the constrained lumping matrices with different \hat{n} . In the simple examples of the present paper, when \hat{n} is large, the resultant lumping matrix coincides with the solution having the global minimum error given by the first optimization approach in Section 3A.

3A. The determination of constrained approximate lumping matrices with global minimum error

In this section we will present an optimization method to determine the constrained lumping matrix with the global minimum error of the invariance of \mathcal{M} to $J^T(y)$. It is not necessary for the new optimization method to solve the matrix equations determining M and consequently to choose an initial value for iteration.

Since we only consider the invariance of \mathcal{M} , when M satisfies the condition

$$MM^T = I_{\hat{n}} \quad (25)$$

the best choice of \bar{M} is M^T (Li and Rabitz, 1990). The approximately lumped system can be described by

$$d\hat{y}/dt = Mf(M^T\hat{y}). \quad (26)$$

In order to determine the approximate lumping matrix we need to minimize the error

$$Z(y) = \text{tr}[E^T(y)E(y)] \quad \forall y \in Y_n \quad (27)$$

where as shown previously

$$E(y) = (I_n - M^T M)J^T(y)M^T. \quad (28)$$

Then we have

$$\begin{aligned} Z(y) &= \text{tr}[E^T(y)E(y)] \\ &= \text{tr}[MJ(y)(I_n - M^T M)(I_n - M^T M)J^T(y)M^T] \\ &= \text{tr}[MJ(y)(I_n - M^T M)J^T(y)M^T]. \end{aligned} \quad (29)$$

Again following the previous work on exact lumping,

$J^T(y)$ can be decomposed into a linear combination of appropriate constant matrices A_k ($k = 1, 2, \dots, m$), i.e.

$$J^T(y) = \sum_{k=1}^m a_k(y)A_k \quad (30)$$

where m is less than n^2 . If y can take any value in the whole Y_n -space, it is reasonable to expect the coefficients $a_k(y)$ to take on any real number, or at least approximately so, and then A_k s can be treated equally without consideration of these coefficients. Thus the determined \mathcal{M} should be as nearly all A_k -invariant as possible, suggesting that the total error Z can be simply defined as

$$Z = \text{tr} \sum_{k=1}^m MA_k^T(I_n - M^T M)A_k M^T. \quad (31)$$

The problem then becomes

$$\begin{aligned} \text{minimize } Z &= \text{tr} \sum_{k=1}^m MA_k^T(I_n - M^T M)A_k M^T \\ \text{subject to } MM^T &= I_{\hat{n}}. \end{aligned} \quad (32)$$

For the constrained lumping problem the lumping matrix M can be represented as

$$M = \begin{pmatrix} M_G \\ M_D \end{pmatrix} \quad (33)$$

where M_G is given and also required to satisfy $M_G M_G^T = I_{\hat{n}-r}$; M_D will be determined and satisfy $M_D M_D^T = I_r$ (where r is the row number of M_D) as well. Then we have

$$\begin{aligned} Z &= \text{tr} \sum_{k=1}^m \begin{pmatrix} M_G \\ M_D \end{pmatrix} A_k^T (I_n - M_G^T M_G - M_D^T M_D) A_k \begin{pmatrix} M_G^T \\ M_D^T \end{pmatrix} \\ &\quad - M_D^T M_D) A_k (M_G^T M_D^T). \end{aligned} \quad (34)$$

Using the property of the trace of a symmetric matrix, eq. (34) can be decomposed as follows:

$$\begin{aligned} Z &= \text{tr} M_D \sum_{k=1}^m A_k^T (I_n - M_G^T M_G - M_D^T M_D) A_k M_D^T \\ &\quad + \text{tr} M_G \sum_{k=1}^m A_k^T (I_n - M_G^T M_G) A_k M_G^T \\ &\quad - \text{tr} M_G \sum_{k=1}^m A_k^T M_D^T M_D A_k M_G^T \\ &= \text{tr} M_D \sum_{k=1}^m A_k^T (I_n - M_G^T M_G - M_D^T M_D) A_k M_D^T \\ &\quad + \text{tr} M_G \sum_{k=1}^m A_k^T (I_n - M_G^T M_G) A_k M_G^T \\ &\quad - \text{tr} M_D \sum_{k=1}^m A_k M_G^T M_G A_k^T M_D^T. \end{aligned} \quad (35)$$

Notice that the three matrices on the right-hand side of eq. (35) are all nonnegative definite. Therefore regardless of the chosen M_D , the first two terms are nonnegative and the last term is nonpositive. This observation suggests finding the M_D such that the last

term has the largest magnitude, thus subtracting from the first two terms as much as possible. It is well known that a symmetric matrix has a full set of orthogonal eigenvectors. Since M_D must satisfy the restriction

$$M_D M_D^T = I_r, \quad (36)$$

the r eigenvectors of the matrix $\sum_{k=1}^m A_k M_G^T M_G A_k^T$ with the largest sum of their eigenvalues solve the problem posed above and the sum is just the magnitude of the last term in eq. (35) (Golub, 1970). Meanwhile, M_D must satisfy another restriction:

$$M_D M_G^T = 0. \quad (37)$$

This restriction can be realized from determination of the eigenvalues and eigenvectors of the matrix

$$Y(1) = \sum_{k=1}^m A_k M_G^T M_G A_k^T + c M_G^T M_G \quad (38)$$

where c is a positive constant. Since the columns of M_G^T are eigenvectors of the matrix $c M_G^T M_G$, when c is large enough and the eigenvectors are arranged in the nonincreasing order of their eigenvalues, the first $\hat{n} - r$ eigenvectors of $Y(1)$ can be as close as possible to M_G^T and the other eigenvectors are orthogonal to it. Therefore, the latter r eigenvectors of $Y(1)$ are a good choice to represent M_D^T , because the result gives the largest magnitude of the last term in eq. (35) under the restriction of eq. (37). However, this choice of M_D will not definitely give the smallest values of the first two terms in eq. (35) and consequently Z . Considering that M_D needs to satisfy eq. (37), then each row of M_D must be a linear combination of the last $n - \hat{n} + r$ eigenvectors of $Y(1)$. Let these $n - \hat{n} + r$ eigenvectors compose the matrix X . When the eigenvalues of $Y(1)$ differ very much, the best M_D^T , which gives the smallest Z , most probably are linear combinations of the first a few columns of X , because the other columns can only yield a very small value for the last term in eq. (35). Let

$$M_D^T = XP \quad (39)$$

where P is an $[(n - \hat{n} + r) \times r]$ -matrix. Taking account of eq. (36) we obtain

$$M_D M_D^T = P^T X^T X P = P^T P = I_r. \quad (40)$$

This implies that all columns of P are orthogonal and normalized. Hence, the magnitude of each element of P is equal to or less than unity, which simplifies the determination of it. Using any of a variety of available programs (say, the IMSL routine ZXMW for determining the global minimum with the presence of constraints) the resultant X will determine P and consequently M_D .

In practice we cannot directly use eqs (34) and (38) to determine Z and $Y(1)$. This comment follows owing to the nonexactness of eq. (28) to describe the deviation of the invariance of \mathcal{M} to $J^T(y)$. The exact determination of the deviation requires the concept of the degree of coincidence of two subspaces defined in

our previous paper (Li and Rabitz, 1990), i.e. the degree of coincidence of \mathcal{M} and the image of it upon $J^T(y)$. According to the definition of the degree of coincidence, each subspace must have an orthonormal basis. Therefore, if we use eq. (28) to represent the deviation of the invariance of \mathcal{M} to $J^T(y)$, we need to transform the matrix $J^T(y) M^T$ to an orthogonal one. When we use eq. (34) to describe Z , we also need to orthonormalize the matrix $A_k (M_G^T M_D^T)$. Similarly, when we determine $Y(1)$, we need to orthonormalize the matrix $A_k M_G^T$. Let $Q_{(k)}^T$, $Q(G)_{(k)}^T$ and $Q(D)_{(k)}^T$ represent the orthonormalized matrices $A_k M^T$, $A_k M_G^T$ and $A_k M_D^T$, respectively. Then eqs (31) and (34) can be revised as

$$Z = \text{tr} \sum_{k=1}^m Q_{(k)} (I_n - M^T M) Q_{(k)}^T \quad (41)$$

$$Z = \text{tr} \sum_{k=1}^m Q(G)_{(k)} (I_n - M_G^T M_G - M_D^T M_D) Q(G)_{(k)}^T + \text{tr} \sum_{k=1}^m Q(D)_{(k)} (I_n - M_G^T M_G - M_D^T M_D) Q(D)_{(k)}^T. \quad (42)$$

Similarly, in eq. (38) we need to orthonormalize $A_k M_G^T$:

$$Y(1) = \sum_{k=1}^m Q(G)_{(k)}^T Q(G)_{(k)} + c M_G^T M_G. \quad (43)$$

It is well known that the minimal A_k -invariant subspace \mathcal{M} over a given subspace \mathcal{M}_G coincides with $\sum_{j=1}^{s-1} A_k^j \mathcal{M}_G$, where s is the rank of A_k , and $A_k^0 = I_n$ (Gohberg *et al.*, 1986). Equation (43) only contains \mathcal{M}_G and $A_k \mathcal{M}_G$, so it does not give the whole picture of the invariance of \mathcal{M} to all A_k . When eq. (43) is used to determine M_D with higher r , the solution probably is certain linear combinations of all columns of X . This comment arises because in this case the first few columns do not span the smallest simultaneously all A_k -invariant subspace over \mathcal{M}_G . The details can be found in Section 3B. If this happens, then this approach will lose its advantage. In order to overcome this problem, one can determine M_D from lower r to higher r in a step-wise fashion. After M_D at lower r has been obtained, one can use $(M_G^T | M_D^T)$ to construct the first term of eq. (43) again. This just adds terms of $A_k^j \mathcal{M}_G$. Then M_D at higher r can be determined by the new $Y(1)$.

After the determination of the eigenvector matrix $R(1)$ of $Y(1)$, we can use the IMSL routine ZXMW to determine P and consequently M_D by minimization of Z under the constraint that the elements $|P_{ij}| \leq 1$. This optimization approach does not need the initial values for P_{ij} . Therefore, in principle, it can be used for lumping problems for any dimension. However, only when the number of the unknown parameters to be determined is not large can ZXMW reach the global minimum solution. Otherwise the solution may possess local minima even if the range $|P_{ij}| \leq 1$ appears small. In Section 3B we will present a direct

approach to circumvent this problem. The solutions of the direct approach are the same or close to those given by the above optimization method with the global minimum error. Therefore, the results obtained by the direct approach can be used to diminish the region of search for the optimization. This overall approach combining the methods of Sections 3A and 3B will be illustrated by the examples used in our previous paper.

3B. The direct determination of constrained approximate lumping matrices

Considering the difficulty reaching the global minimum solution with the above optimization approach and also that some amount of error is acceptable in practice, it would be desirable to develop a direct way for determining the constrained approximate lumping schemes with satisfactory accuracy. Using the concept of the minimal A_k -invariant subspace \mathcal{M} over a given subspace \mathcal{M}_G , we have built such an approach described below.

It is well known that the minimal invariant subspace \mathcal{M} for an $(n \times n)$ -matrix A over a given subspace $\text{Im } B$ coincides with

$$\mathcal{M} = \sum_{j=0}^{\infty} \text{Im}(A^j B) = \sum_{j=0}^{s-1} \text{Im}(A^j B) \quad (44)$$

for every integer s greater than or equal to the rank or the degree of a minimal polynomial for A [in particular, $\mathcal{M} = \sum_{j=0}^{n-1} \text{Im}(A^j B)$] (Gohberg *et al.*, 1986). We know that

$$\sum_{j=0}^{s-1} \text{Im}(A^j B) = \text{Im}(B \ AB \ \dots \ A^{s-1} B) \quad (45)$$

and the orthogonal decomposition of the n -dimensional real space \mathcal{R}^n is

$$\mathcal{R}^n = \text{Im}(B \ AB \ \dots \ A^{s-1} B) \oplus \text{Ker} \begin{bmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{s-1} \end{bmatrix}. \quad (46)$$

In order to determine $\text{Im}(B \ AB \ \dots \ A^{s-1} B)$ we can first determine the kernel by solving the following equation:

$$\begin{bmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{s-1} \end{bmatrix} X = 0. \quad (47)$$

Suppose the dimension of $\text{Im } X$ is $n-l$. After the determination of X then the matrix representation M^T of the smallest A -invariant subspace \mathcal{M} with dimension l over $\text{Im } B$ can be determined by solving the equation

$$X^T M^T = 0. \quad (48)$$

It is straightforward to determine the minimal simultaneously A_k ($k = 1, 2, \dots, m$)-invariant sub-

space \mathcal{M} over the subspace $\text{Im } B$. We only need to determine X first by solving the following equation:

$$\begin{bmatrix} B^T \\ B^T A_1^T \\ \vdots \\ B^T (A_1^T)^{s_1-1} \\ \vdots \\ B^T \\ B^T A_m^T \\ \vdots \\ B^T (A_m^T)^{s_m-1} \end{bmatrix} X = 0 \quad (49)$$

where s_k ($k = 1, \dots, m$) is greater than or equal to the rank of A_k , and then solving eq. (48) to determine M . In the current problem $B = M_G^T$, $\mathcal{R}^n = Y_n$ and the resultant M is the exact lumping matrix containing M_G with the smallest row number l .

When we want to proceed further to find good-quality approximate lumping matrices with \hat{n} less than l , we need first to determine higher-dimensional $\text{Im } X$ which are as nearly as possible orthogonal to

$$\begin{bmatrix} M_G \\ M_G A_1^T \\ \vdots \\ M_G (A_1^T)^{s_1-1} \\ \vdots \\ M_G \\ M_G A_m^T \\ \vdots \\ M_G (A_m^T)^{s_m-1} \end{bmatrix} \quad (50)$$

Then the resultant \mathcal{M} will be as nearly all A_k -invariant as possible. The corresponding M s are good approximate lumping matrices containing M_G with \hat{n} less than l . This consideration is equivalent to finding the subspace $\text{Im } X$, which is simultaneously as nearly orthogonal to $\text{Im } M_G^T$, $\text{Im} (M_G A_1^T)^T, \dots$, $\text{Im} [M_G (A_1^T)^{s_1-1}]^T$, $\text{Im } M_G^T$, $\text{Im} (M_G A_m^T)^T, \dots$, $\text{Im} [M_G (A_m^T)^{s_m-1}]^T$ as possible. This X can be readily determined by using the concept of the degree of coincidence between two subspaces given in our previous paper (Li and Rabitz, 1990).

Let $Q(G)_{(ki)}^T$ ($k = 1, 2, \dots, m; i = 0, 1, \dots, s_k - 1$) be the orthonormal matrix representation of $\text{Im} [M_G (A_k^T)^i]^T$ using the Schmidt orthogonalization method one can transform $[M_G (A_k^T)^i]^T$ to $Q(G)_{(ki)}^T$. First we define a matrix

$$Y(2) = \sum_{k=1}^m \sum_{i=0}^{s_k-1} Q(G)_{(ki)}^T Q(G)_{(ki)}. \quad (51)$$

If we choose a set of orthonormal basis for $\text{Im } X$, i.e.

$$X^T X = I_{n-\hat{n}} \quad (52)$$

then the problem becomes the determination of X , which gives the smallest trace

$$\min_{X^T X = I_{n-\hat{n}}} \text{tr } X^T Y(2) X. \quad (53)$$

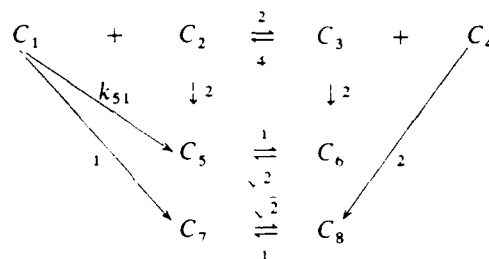
The solution can be readily obtained by determining the eigenvalues and eigenvectors of $Y(2)$ (Bellman, 1970). The $n - \hat{n}$ eigenvectors with the smallest sum of their eigenvalues are X and the rest of the eigenvectors compose M^T . When all the eigenvalues are distinct, the solution for M with a specified \hat{n} is unique. If there exist multiple eigenvalues, the sets of eigenvectors with the same sum of eigenvalues are all solutions. When the eigenvectors of $Y(2)$ are arranged according to the nonincreasing order of their eigenvalues, the last $n - \hat{n}$ eigenvectors are X and the first \hat{n} eigenvectors are M^T . Therefore, the eigenvector matrix $R(2)$ of $Y(2)$ supplies all lumping matrices with different \hat{n} .

There are two further issues we need to consider. First, sometimes $M_G A_i^T$ is a null matrix. In this case the contribution of A_i to the determination of the lumping matrix can be neglected. In order to avoid this situation, we can use the resultant M from other A_k with row number 1 higher than M_G as a new M_G to calculate $M_G A_i^T$. If $M_G A_i^T$ for the new M_G is still a null matrix, we can use the resultant M with row number 2 higher than the original M_G as a new M_G to calculate $M_G A_i^T$ and so on. Second, as in the discussion in Section 3A, in order to satisfactorily assure that the resultant M_D is orthogonal to M_G , one can multiply M_G in eq. (50) by a large positive constant c .

Notice that the M_D obtained by eq. (53) will not definitely give the minimum Z . As shown below in the simple examples, when \hat{n} is close to the dimension of the smallest simultaneously A_k -invariant subspace over M_G , the solutions of this direct approach really have the global minimum Z . In other cases, however, the solutions of the direct approach are still close to the global minimum ones. Therefore, we can readily determine the best lumping matrices with large \hat{n} by the direct approach. For the lumping matrices with small \hat{n} , if the errors of the solutions obtained by the direct approach are acceptable, one can directly use the resultant M . Otherwise, one can use the optim-

4. EXAMPLES

The methods proposed in this paper will be illustrated by the following reaction scheme, where the C_i s are species and the numbers are unitless rate constants:



When $k_{51} = 1$, this mechanism admits some exact lumping solutions. By changing the rate constant k_{51} to 0.9 (example 1) and 0.1 (example 2) the system contains some exact and approximate lumping schemes.

Letting y_i represent the concentration of C_i , it is easy to write out the kinetic equations and the transpose of the corresponding Jacobian matrix $J^T(y)$:

$$\begin{aligned}
 dy_1/dt &= -(1 + k_{51})y_1 - 2y_1y_2 + 4y_3y_4 \\
 dy_2/dt &= -2y_2 - 2y_1y_2 + 4y_3y_4 \\
 dy_3/dt &= -2y_3 - 4y_3y_4 + 2y_1y_2 \\
 dy_4/dt &= -2y_4 - 4y_3y_4 + 2y_1y_2 \\
 dy_5/dt &= -y_5 + k_{51}y_1 + 2y_2 + \sqrt{2}y_6 \\
 dy_6/dt &= -\sqrt{2}y_6 + 2y_3 + y_5 \\
 dy_7/dt &= -\sqrt{2}y_7 + y_1 + y_8 \\
 dy_8/dt &= -y_8 + 2y_4 + \sqrt{2}y_7
 \end{aligned} \tag{54}$$

$$J^T(y) = \begin{bmatrix}
 -2y_2 - 1 - k_{51} & -2y_2 & 2y_2 & 2y_2 & k_{51} & 0 & 1 & 0 \\
 -2y_1 & -2(1 + y_1) & 2y_1 & 2y_1 & 2 & 0 & 0 & 0 \\
 4y_4 & 4y_4 & -2(1 + 2y_4) & -4y_4 & 0 & 2 & 0 & 0 \\
 4y_5 & 4y_3 & -4y_3 & -2(1 + 2y_3) & 0 & 0 & 0 & 2 \\
 & 0 & & & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\
 & & & & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 \\
 & & & & 0 & 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\
 & & & & 0 & 0 & 1 & -1
 \end{bmatrix}$$

$J^T(y)$ can be represented as

$$J^T(y) = A_0 + \sum_{k=1}^4 y_k A_k \tag{55}$$

ization method given in Section 3A to determine M and the results of the direct approach may be used to diminish the region of the unknown parameters.

where

$$A_0 = \begin{pmatrix} -1-k_{s1} & 0 & 0 & 0 & k_{s1} & 0 & 1 & 0 \\ 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2 \\ & & & & -1 & 1 & 0 & 0 \\ & & & 0 & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ & & & & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ & & & 0 \\ & & & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -2 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ & & & 0 & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 4 & -4 & -4 \\ & & & 0 \\ & & & 0 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 4 & 4 & -4 & -4 \\ 0 & 0 & 0 & 0 \\ & & & 0 & 0 \end{pmatrix}$$

This information will be used in the examples below.

4A. Example 1

We will first employ the optimization approach presented in Section 3A to determine the constrained lumping matrices with the global minimum error. Then the direct approach given in Section 3B will be

employed. The results obtained by these approaches will be compared with each other.

Let $k_{s1} = 0.9$ and the given part of the lumping matrix is taken as

$$M_G = (0.0000 \ 0.0000 \ 0.0000 \ 0.0000 \ 0.5000 \ 0.5000 \ 0.5000 \ 0.5000).$$

Utilizing eq. (43) and letting $c = 2$, we obtain the symmetric matrix

$$Y(1) = \begin{pmatrix} 0.2313 & 0.2434 & 0.2434 & 0.2434 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2434 & 0.2562 & 0.2562 & 0.2562 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2434 & 0.2562 & 0.2562 & 0.2562 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2434 & 0.2562 & 0.2562 & 0.2562 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \end{pmatrix}$$

The eigenvalues and corresponding eigenvectors are given as follows:

$$\lambda_i = \begin{matrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$R(1) = \begin{pmatrix} 0.0000 & 0.4809 & 0.8768 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5062 & -0.2776 & -0.1538 & 0.8019 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5062 & -0.2776 & 0.7713 & -0.2678 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5062 & -0.2776 & -0.6176 & -0.5341 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0846 & 0.7071 & -0.4928 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0846 & -0.7071 & -0.4928 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.7815 & 0.0000 & 0.3732 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.6124 & 0.0000 & 0.6124 \end{pmatrix}$$

This example is very special. No matter what value of c we choose M_G^T is an eigenvector of $Y(1)$. For other problems c should be big enough to guarantee that each column of M_G^T is an eigenvector of $Y(1)$. However, in the present example c must be larger than 1. Otherwise, the eigenvalue of M_G^T for $Y(1)$ is not larger than 1 and M_G^T cannot be located in the first column in $R(1)$.

Since the first column of $R(1)$ is M_G^T and other columns are orthogonal to it, any row of M_D must be a certain linear combination of these seven columns, which compose the matrix X . One can see that only

the second column of $R(1)$ in X has a nonzero eigenvalue. If we want to determine M_D with $r = 1$, this column most probably is the solution owing to its giving the largest magnitude 1 to the last term in eq. (35). Indeed, using the IMSL routine ZXWWD we find the global minimum solution of the linear combination coefficient vector

$$P = (1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0)^T$$

and this corresponds to M_D^T being the second column of $R(1)$. Then the resultant best lumping matrix with $\hat{n} = 2$ is

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4809 & 0.5062 & 0.5062 & 0.5062 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

This lumping matrix M may now be used as M_G to construct the first term of eq. (43) again for the determination of the lumping matrix with $\hat{n} = 3$. The resultant $Y(1)$ and $R(1)$ are the following:

$$Y(1) = \begin{pmatrix} 2.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.3333 & 0.3333 & 0.3333 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3333 & 1.3333 & 0.3333 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.3333 & 0.3333 & 1.3333 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \end{pmatrix}$$

$$\lambda_i = \begin{matrix} 5 & 2 & 2 & 1 & 1 & 0 & 0 & 0 \end{matrix}$$

$$R(1) = \begin{pmatrix} 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5774 & 0.0000 & -0.8165 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5774 & 0.7071 & 0.4083 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5774 & -0.7071 & 0.4083 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0846 & 0.7071 & -0.4928 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.0846 & -0.7071 & -0.4928 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.7815 & 0.0000 & 0.3732 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.6124 & 0.0000 & 0.6124 \end{pmatrix}$$

In order to locate M_G^T in the first column of the new $R(1)$ we choose $c = 5$. We find the first and the second rows of M_D simultaneously by the determination of the (7×2) -matrix P . The result is

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T$$

Following the same procedure we use this M as M_G to construct a new $Y(1)$ and determine the best lumping matrix with $\hat{n} = 4$. The resultant $Y(1)$ and $R(1)$ are the same. In this case we found that the solution is not unique. For example, the following two lumping matrices have the same total error:

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5774 & 0.5774 & 0.5774 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5774 & 0.5774 & 0.5774 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.8165 & 0.4083 & 0.4083 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

The resultant best lumping matrix with $\hat{n} = 3$ is

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5774 & 0.5774 & 0.5774 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

Any linear combination of the last rows in the two matrices (provided it is normalized) can be used as the new last row to give a lumping matrix with the same accuracy. For example, we have

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5774 & 0.5774 & 0.5774 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.7071 & 0.0000 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

When we use columns 1–5 of $R(1)$ to construct the lumping matrix with $\hat{n} = 5$, it is an exact one. This is equivalent to the following simple lumping matrix:

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

These resultant lumping matrices are similar to the following ones obtained by solving the matrix equations in our previous paper (Li and Rabitz, 1990), except for $\hat{n} = 3$:

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4843 & 0.5101 & 0.5026 & 0.5026 & -0.0012 & 0.0040 & -0.0072 & 0.0044 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4839 & 0.5098 & 0.5029 & 0.5029 & -0.0013 & 0.0040 & -0.0073 & 0.0047 \\ 0.0000 & 0.0000 & 0.7071 & -0.7071 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5211 & 0.4721 & 0.4913 & 0.5139 & -0.0052 & 0.0051 & -0.0030 & 0.0031 \\ 0.0338 & -0.0186 & 0.7137 & -0.6994 & -0.0002 & 0.0002 & -0.0001 & 0.0001 \\ -0.6934 & 0.7188 & 0.0484 & -0.0033 & 0.0055 & -0.0076 & 0.0049 & -0.0028 \end{pmatrix}$$

A lumping matrix M can be considered as the matrix representation of a subspace. Then the similarity of the lumping matrices given by the present approach and the original one may be determined by the corresponding degree of coincidence between the two subspaces, d_c . For $\hat{n} = 2, 3$ and 4 , we have $d_c = 0.99, 0.67$ and 0.92 , respectively. They are very close for $\hat{n} = 2$ and 4 . In our previous paper, we used eq. (28) to describe the deviation of the invariance of \mathcal{M} to $J^T(y)$. Hence, the results have a larger error. For $\hat{n} = 3$ the lumping matrix obtained by our previous paper is a local minimum solution, which can also be obtained by the present optimization approach if we constrain the unknown parameters in the suitable region. From our previous paper one can find that the initial values of iteration we chose for the matrix equations did not contain one which is near the global minimum solution and then we failed to find it.

Utilizing eq. (21) and the present optimization approach, we obtain the lumped kinetic equations for the new lumping matrices validated in the whole Y_n space as follows:

Lumped kinetic equations with $\hat{n} = 2$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.9755\hat{y}_2 \\ d\hat{y}_2/dt &= -1.9768\hat{y}_2 - 0.01361\hat{y}_2^2 \end{aligned} \quad (56)$$

Lumped kinetic equations with $\hat{n} = 3$:

$$\begin{aligned} d\hat{y}_1/dt &= 0.95\hat{y}_2 + 1.7321\hat{y}_3 \\ d\hat{y}_2/dt &= -1.9000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.3333\hat{y}_3^2 \\ d\hat{y}_3/dt &= -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.7698\hat{y}_3^2 \end{aligned} \quad (57)$$

Lumped kinetic equations with $\hat{n} = 4$ (three equivalent lumped models):

$$\begin{aligned} d\hat{y}_1/dt &= 0.9500\hat{y}_2 + 1.7321\hat{y}_3 \\ d\hat{y}_2/dt &= -1.9000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.3333\hat{y}_3^2 \\ &\quad - 2.0000\hat{y}_4^2 \\ d\hat{y}_3/dt &= -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.7698\hat{y}_3^2 \\ &\quad + 1.1547\hat{y}_4^2 \end{aligned} \quad (58)$$

$$d\hat{y}_4/dt = -2.0000\hat{y}_4$$

$$d\hat{y}_1/dt = 0.9500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.9000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.6330\hat{y}_2\hat{y}_4 \\ + 1.8856\hat{y}_3\hat{y}_4 + 1.3333\hat{y}_3^2 + 0.6667\hat{y}_4^2$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.9428\hat{y}_2\hat{y}_4 \\ - 1.0887\hat{y}_3\hat{y}_4 - 0.7698\hat{y}_3^2 - 0.3849\hat{y}_4^2$$

(59)

$$d\hat{y}_4/dt = -2.0000\hat{y}_4 + 1.8856\hat{y}_2\hat{y}_3 - 2.6667\hat{y}_2\hat{y}_4 \\ - 3.0792\hat{y}_3\hat{y}_4 - 2.1773\hat{y}_3^2 - 1.0887\hat{y}_4^2$$

$$d\hat{y}_1/dt = 0.9500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.9000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 - 1.4142\hat{y}_2\hat{y}_4 \\ - 1.6330\hat{y}_3\hat{y}_4 + 1.3333\hat{y}_3^2$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 + 0.8165\hat{y}_2\hat{y}_4 \\ + 0.9430\hat{y}_3\hat{y}_4 - 0.7698\hat{y}_3^2 \quad (60)$$

$$d\hat{y}_4/dt = -2.0000\hat{y}_4 + 1.6330\hat{y}_2\hat{y}_3 - 2.0000\hat{y}_2\hat{y}_4 \\ - 2.3094\hat{y}_3\hat{y}_4 + 1.8856\hat{y}_3^2$$

For comparison the solutions of \hat{y}_i (other lumped species \hat{y}_i have the same accuracy as that of \hat{y}_1) of eqs (54) (original model) and (56)–(58) (approximately lumped models) for different initial values are given in Figs 1–3. Equations (58)–(60) have the same accuracy. The results are very satisfactory for all chosen initial conditions, even if $\hat{n} = 2$. The differences between the present lumping matrices and those obtained in our previous paper are not very large, but the accuracy of the new lumping matrices is much higher.

Now we apply the second approach in Section 3B to determine the approximate lumping matrices directly. Using eqs (50) and (51) one can obtain matrix $Y(2)$. Since A_0 has the highest rank, 6, we simply take all $s_k = 6$. The resultant $Y(2)$ and its eigenvalues and eigenvector matrix $R(2)$ are given below:

$$Y(2) = \begin{pmatrix} 0.9891 & 1.1469 & 1.1469 & 1.1469 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 1.3370 & 1.3370 & 1.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 1.3370 & 1.3370 & 1.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 1.3370 & 1.3370 & 1.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.2500 & 1.2500 & 1.2500 & 1.2500 \end{pmatrix}$$

$$R(2) = \begin{pmatrix} \lambda_i = 5.0000 & 4.9959 & 0.0041 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.4442 & 0.8959 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & 0.0000 & -0.8165 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & 0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & -0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & 0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & -0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5443 & 0.0000 & 0.6736 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.8165 & 0.0000 & 0.2887 \end{pmatrix}$$

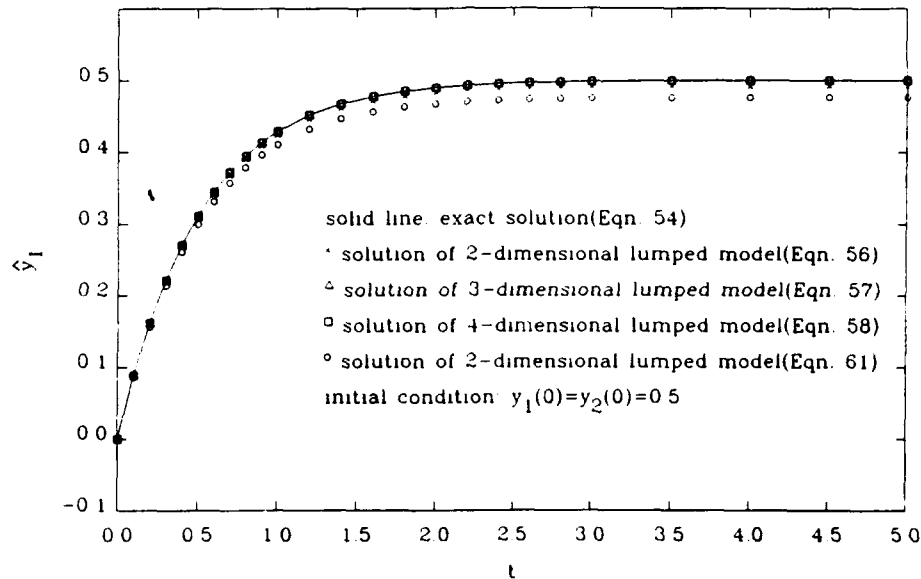


Fig. 1. Comparison between the solutions of \hat{y}_1 for eqs (54), (56)–(58) and (61) [initial condition: $y_1(0) = y_2(0) = 0.5$, others are zero].

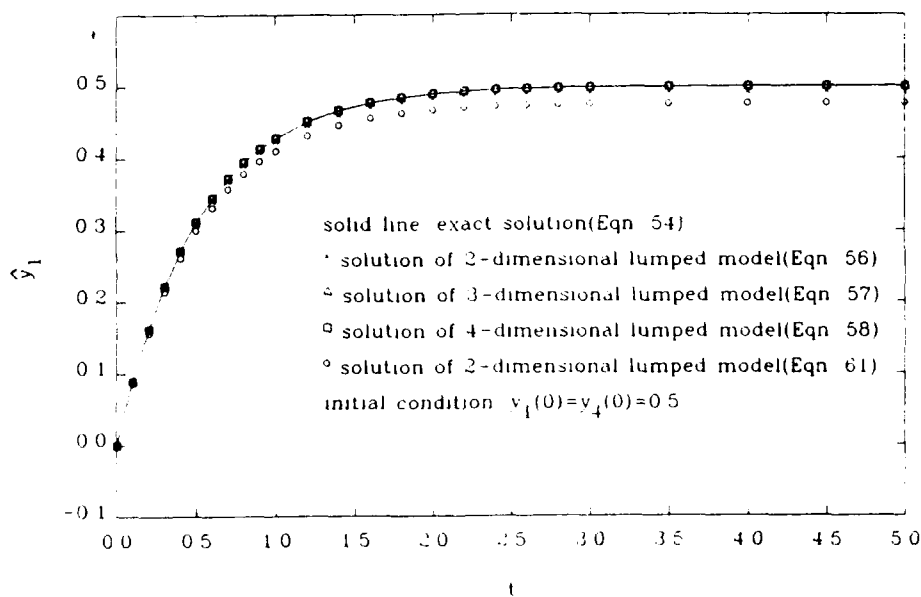


Fig. 2. Comparison between the solutions of \hat{y}_1 for eqs (54), (56)–(58) and (61) [initial condition: $y_1(0) = y_4(0) = 0.5$, others are zero].

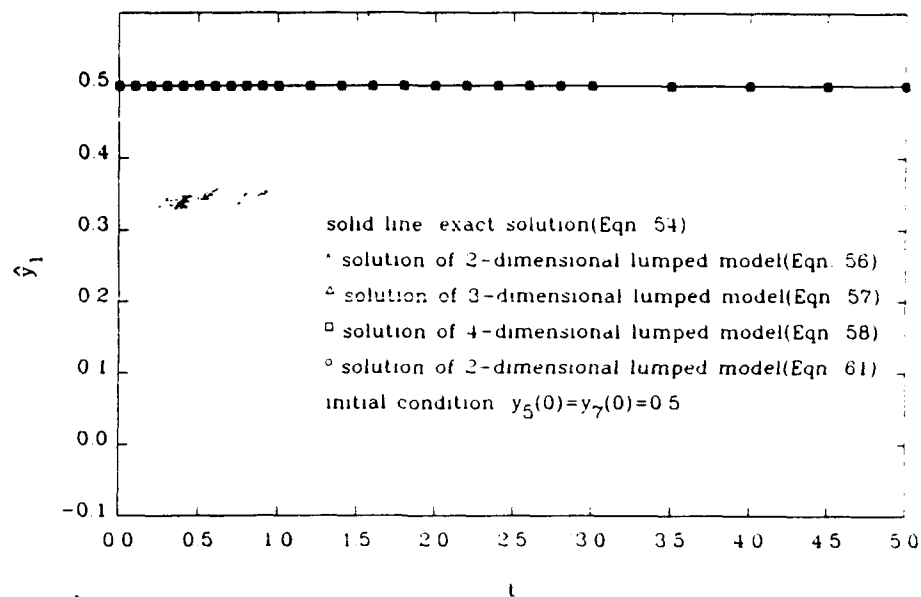


Fig. 3. Comparison between the solutions of \hat{y}_1 for eqs (54), (56)–(58) and (61) [initial condition: $y_5(0) = y_7(0) = 0.5$, others are zero].

In this example, $M_G A_i^T = 0$ ($i = 1-4$). Therefore, we use the first two columns of $R(2)$ to calculate $M_G A_i^T$ ($i = 1-4$) again. In order to force the M_G^T to be the first column of $R(2)$ we multiply M_G by 2. The resultant new $Y(2)$ and $R(2)$ with the corresponding eigenvalues are the following:

However, the fourth and fifth eigenvalues are equal, and the best lumping matrix with $\hat{n} = 4$ is not unique. The first three columns of $R(2)$ with either one of the columns 4 and 5 or any linear combination of these two columns (provided the resultant vector is nor-

$$Y(2) = \begin{pmatrix} 5.9891 & 1.1469 & 1.1469 & 1.1469 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 6.3370 & 1.3370 & 1.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 1.3370 & 6.3370 & 1.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.1469 & 1.3370 & 1.3370 & 6.3370 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \end{pmatrix}$$

$$\lambda_i = 10.0000 \quad 9.9959 \quad 5.0042 \quad 5.0000 \quad 5.0000 \quad 0.0000 \quad 0.0000 \quad 0.0000$$

$$R(2) = \begin{pmatrix} 0.0000 & 0.4442 & 0.8959 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & 0.0000 & -0.8165 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & 0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5173 & -0.2565 & -0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & 0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & -0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5443 & 0.0000 & 0.6736 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.8165 & 0.0000 & 0.2887 \end{pmatrix}$$

The resultant $R(2)$ is the same, but the eigenvalues are different. According to the second approach the first two columns of $R(2)$ form the best lumping matrix with $\hat{n} = 2$, the first three columns of $R(2)$ form the best lumping matrix with $\hat{n} = 3$. Since the eigenvalues of the first three eigenvectors are distinct, the best lumping matrices with $\hat{n} = 2$ and 3 are unique.

malized) will give lumping matrices having the same accuracy. The first five columns of $R(2)$ form an exact lumping matrix because the rest of eigenvalues are all zero.

Since M is only a matrix representation of a subspace, row elementary operations (multiply one row

by a constant, interchange the positions of two rows, subtract one row multiplied by a constant from another row) will give another matrix representation of the same subspace (Lang, 1986). These two M s are equivalent. Using the row elementary operations on columns 2 and 3 of $R(2)$ the best lumping matrix with $\hat{n} = 3$ can be represented as

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5774 & 0.5774 & 0.5774 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}.$$

Comparing the results of the two approaches, one can see that the resultant best lumping matrices are the same except for $\hat{n} = 2$. The best lumping matrix with $\hat{n} = 2$ given by the second direct approach is the following:

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.4442 & 0.5173 & 0.5173 & 0.5173 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}.$$

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.3027 & 0.5503 & 0.5503 & 0.5503 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}.$$

The corresponding lumped kinetic equations are as follows:

$$\begin{aligned} d\hat{y}_1/dt &= 1.9739\hat{y}_2 \\ d\hat{y}_2/dt &= -1.9805\hat{y}_2 - 0.0447\hat{y}_2^2. \end{aligned} \quad (61)$$

For comparison the solutions of \hat{y}_1 of eq. (61) for

$$\begin{aligned} M &= \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.2945 & 0.6025 & 0.5220 & 0.5222 & -0.0017 & 0.0271 & -0.0577 & 0.0324 \end{pmatrix} \\ M &= \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.3196 & 0.5953 & 0.5205 & 0.5199 & -0.0297 & 0.0259 & -0.0163 & 0.0201 \\ 0.8486 & -0.5237 & 0.0427 & 0.0304 & -0.0315 & 0.0301 & -0.0224 & 0.0238 \end{pmatrix} \\ M &= \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.5389 & 0.4324 & 0.5427 & 0.4750 & -0.0334 & 0.0275 & -0.0130 & 0.0189 \\ 0.5304 & -0.4455 & 0.3710 & -0.6182 & 0.0080 & -0.0149 & 0.0101 & -0.0031 \\ 0.5537 & -0.4135 & -0.5877 & 0.4207 & 0.0029 & -0.0091 & 0.0069 & -0.0007 \end{pmatrix}. \end{aligned}$$

different initial values are also given in Figs 1–3. The results are quite satisfactory for all chosen initial conditions, but of somewhat lesser quality than in eq. (56) with the first optimization method.

From the comparison of the results for these two approaches one finds that the global minimum solutions of the constrained lumping matrices can be readily obtained by the second direct approach if \hat{n} is close to the dimension of the smallest simultaneously A_k -invariant subspace over \mathcal{M}_G . In other cases the resultant lumping matrix given by the second direct approach is still very close to the global minimum solution. Therefore, taking advantage of this situation one can directly determine the elements of M with

small \hat{n} instead of P_{ij} by the first optimization approach and constrain the region of M_{ij} around the solution given by the second direct approach. From the above example, one can see that the global minimum solutions of M with small \hat{n} are easy to reach in this way.

4B. Example 2

The second example is the same system except that $k_{s1} = 0.1$. For the same M_G as that of example 1 the first approach gives the same best lumping matrices for different \hat{n} as those of example 1, except that $\hat{n} = 2$

These resultant lumping matrices are similar to those lumping matrices obtained by solving the matrix equations in our previous paper (Li and Rabitz, 1990). For comparison those lumping matrices are listed below:

The degree of coincidence between the subspaces corresponding to the present and the original solutions of M , $d_c = 0.98, 0.93$ and 0.96 for $\hat{n} = 2, 3$ and 4 , respectively.

Utilizing eq. (21), the resultant lumped kinetic equations for the new lumping matrices given by the optimization approach validated in the whole Y_n -space are as follows:

Lumped kinetic equations with $\hat{n} = 2$:

$$\begin{aligned} d\hat{y}_1/dt &= 1.8173\hat{y}_2 \\ d\hat{y}_2/dt &= -1.175\hat{y}_2 - 0.2174\hat{y}_2^2 \end{aligned} \quad (62)$$

Lumped kinetic equations with $\hat{n} = 3$:

$$d\hat{y}_1/dt = 0.5500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.1000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.3333\hat{y}_3^2 \quad (63)$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.7698\hat{y}_3^2$$

Lumped kinetic equations with $\hat{n} \leq 4$ (three equivalent lumped models):

$$d\hat{y}_1/dt = 0.5500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.1000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.3333\hat{y}_3^2 - 2.0000\hat{y}_4^2 \quad (64)$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.7698\hat{y}_3^2 + 1.1547\hat{y}_4^2$$

$$d\hat{y}_4/dt = -2.0000\hat{y}_4$$

$$d\hat{y}_1/dt = 0.5500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.1000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 + 1.6330\hat{y}_2\hat{y}_4 + 1.8856\hat{y}_3\hat{y}_4 + 1.3333\hat{y}_3^2 + 0.6667\hat{y}_4^2$$

$$d\hat{y}_4/dt = -2.0000\hat{y}_4 + 1.8856\hat{y}_2\hat{y}_3 - 2.6667\hat{y}_2\hat{y}_4$$

$$-3.0792\hat{y}_3\hat{y}_4 - 1.1773\hat{y}_3^2 - 1.0887\hat{y}_4^2$$

$$d\hat{y}_1/dt = 0.5500\hat{y}_2 + 1.7321\hat{y}_3$$

$$d\hat{y}_2/dt = -1.1000\hat{y}_2 - 1.1547\hat{y}_2\hat{y}_3 - 1.4142\hat{y}_2\hat{y}_4 - 1.6330\hat{y}_3\hat{y}_4 + 1.3333\hat{y}_3^2$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 + 0.8165\hat{y}_2\hat{y}_4 + 0.9430\hat{y}_3\hat{y}_4 - 0.7698\hat{y}_3^2 \quad (66)$$

$$d\hat{y}_4/dt = -2.0000\hat{y}_4 + 1.6330\hat{y}_2\hat{y}_3 - 2.0000\hat{y}_2\hat{y}_4 - 2.3094\hat{y}_3\hat{y}_4 + 1.8856\hat{y}_3^2$$

For comparison the solutions of \hat{y}_1 of eqs (54) (original model) and (62)–(64) (approximately lumped models) for different initial values are given in Figs 4–6. Equations (64)–(66) have the same accuracy. The results are very satisfactory for all chosen initial conditions when $\hat{n} \geq 3$. In contrast, the lumping matrix obtained in our previous paper still has a relatively large error when $\hat{n} = 4$ [see Figs 4 and 6 in Li and Rabitz (1990)].

Similarly, utilizing the second direct approach the matrix $Y(2)$ and its corresponding eigenvalues and eigenvector matrix $R(2)$ are the following:

$$Y(2) = \begin{pmatrix} 5.1340 & 0.3665 & 0.3665 & 0.3665 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3665 & 6.6220 & 1.6220 & 1.6220 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3665 & 1.6220 & 6.6220 & 1.6220 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.3665 & 1.6220 & 1.6220 & 6.6220 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 2.5000 & 2.5000 & 2.5000 & 2.5000 \end{pmatrix}$$

$$\lambda_i = 10.0000 \quad 9.9497 \quad 5.0503 \quad 5.0000 \quad 5.0000 \quad 0.0000 \quad 0.0000 \quad 0.0000$$

$$R(2) = \begin{pmatrix} 0.0000 & 0.1307 & 0.9914 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5724 & -0.0755 & 0.0000 & -0.8165 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5724 & -0.0755 & 0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.5724 & -0.0755 & -0.7071 & 0.4082 & 0.0000 & 0.0000 & 0.0000 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & 0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1361 & -0.7071 & -0.4811 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5443 & 0.0000 & 0.6736 \\ 0.5000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & -0.8165 & 0.0000 & 0.2887 \end{pmatrix}$$

$$d\hat{y}_3/dt = -2.0000\hat{y}_3 + 0.6667\hat{y}_2\hat{y}_3 - 0.9428\hat{y}_2\hat{y}_4 - 1.0887\hat{y}_3\hat{y}_4 - 0.7698\hat{y}_3^2 - 0.3849\hat{y}_4^2 \quad (65)$$

As in example 1 the best lumping matrices with $\hat{n} \geq 3$ obtained by the second direct approach are the same as those given by the first optimization one. The best lumping matrix with $\hat{n} = 2$ given by the second approach is the following:

$$M = \begin{pmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.5000 & 0.5000 & 0.5000 & 0.5000 \\ 0.1307 & 0.5724 & 0.5724 & 0.5724 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{pmatrix}$$

This lumping matrix is still close to the one given by the first approach. The corresponding lumped kinetic equations are given below:

$$\begin{aligned} d\hat{y}_1/dt &= 1.7891\hat{y}_2 \\ d\hat{y}_2/dt &= -1.9846\hat{y}_2 - 0.5129\hat{y}_2^2. \end{aligned} \quad (67)$$

For comparison the solutions of \hat{y}_1 of eq. (67) for different initial values are also given in Figs 4–6. The results are not satisfactory for all chosen initial conditions. However, for $\hat{n} = 2$ the lumping matrix with the global minimum error given by the first optimization approach also has a quite large error.

From these examples one can see that these two approaches are simpler than that given in our previ-

ous paper when applied to determining the lumping schemes validated in the whole composition Y_n -space.

5. CONCLUSION AND DISCUSSION

In the present paper, we have proved that the necessary and sufficient conditions for the existence of exact lumping in the whole composition space become simpler. The invariance of the subspace \mathcal{H} spanned by the row vectors of the lumping matrix M to the transpose of the Jacobian matrix $J^T(y)$ for all values of y in the Y_n -space is sufficient for exact lumping.

A new optimization approach to determine the constrained approximate lumping schemes with the

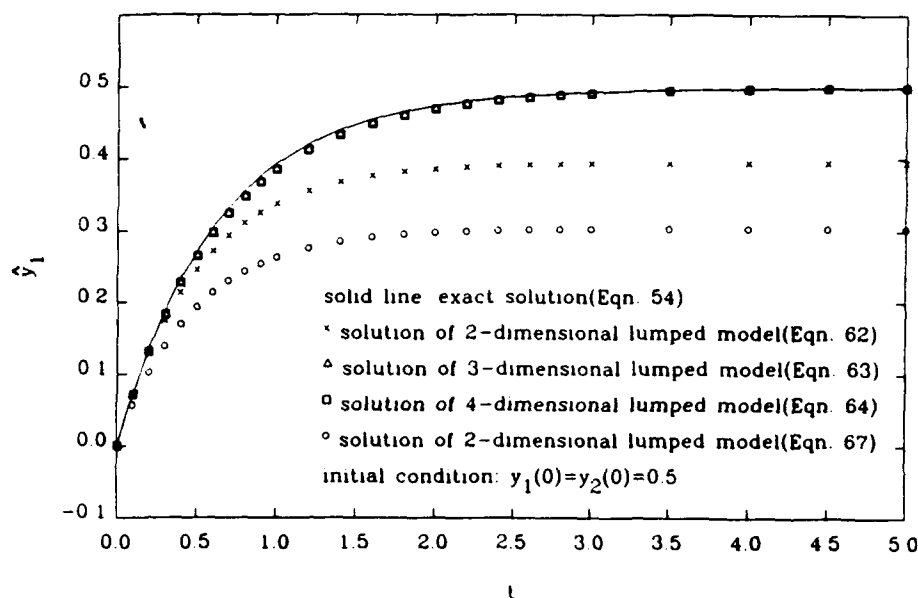


Fig. 4. Comparison between the solutions of \hat{y}_1 for eqs (54), (62)–(64) and (67) [initial condition: $y_1(0) = y_2(0) = 0.5$, others are zero].

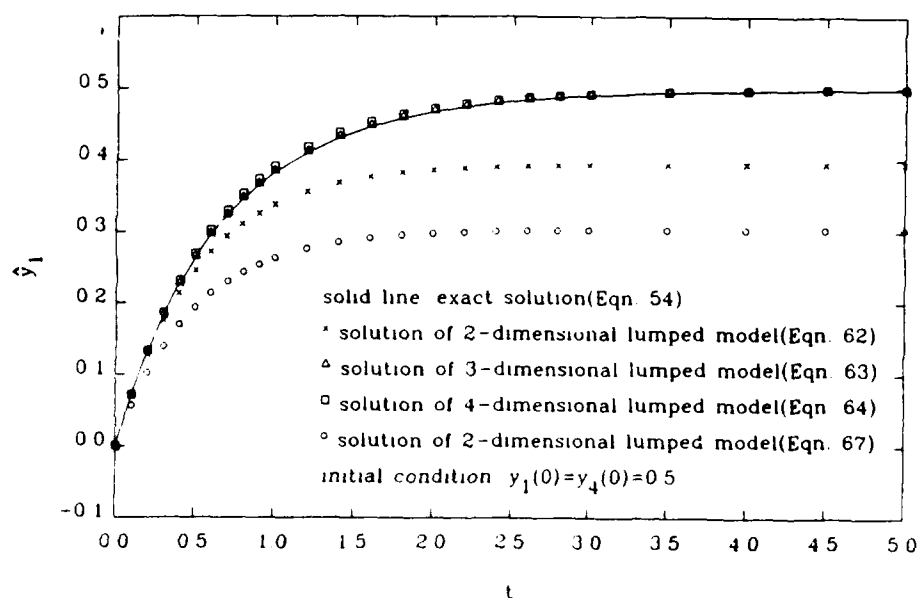


Fig. 5. Comparison between the solutions of \hat{y}_1 for eqs (54), (62)–(64) and (67) [initial condition: $y_1(0) = y_4(0) = 0.5$, others are zero].

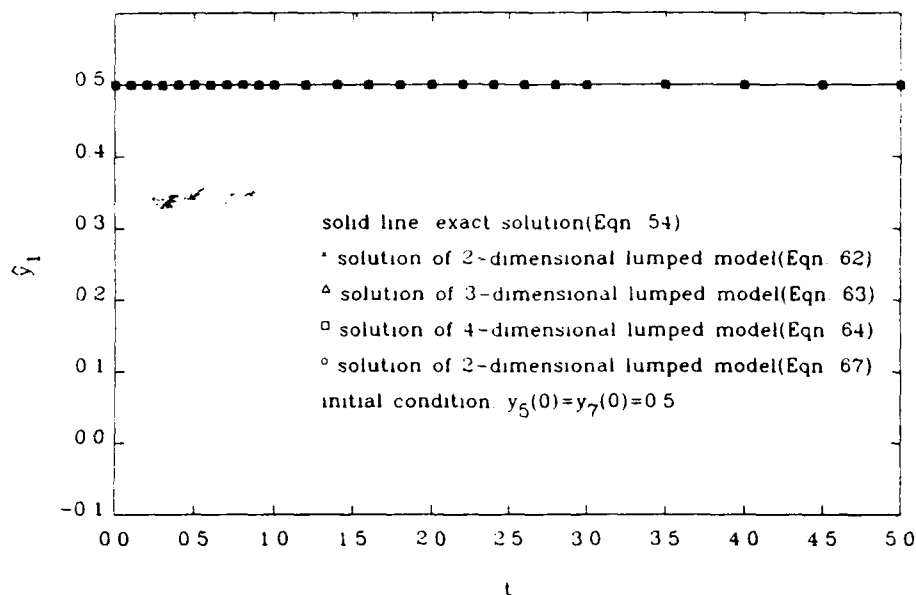


Fig. 6. Comparison between the solutions of \hat{y}_1 for eqs (54), (62)–(64) and (67) [initial condition: $y_5(0) = y_7(0) = 0.5$, others are zero].

global minimum error is presented. This approach is based on the decomposition of the total error. When the approximate lumping schemes are validated in the whole Y_n -space, we can effectively treat all A_k equally. This simplifies the determination of the constrained lumping schemes. Using M_G^T and all orthonormalized $A_k M_G^T$ one can construct a special symmetric matrix $Y(1)$. The rows of the part to be determined M_D of M are linear combinations of those eigenvectors of $Y(1)$ with the largest eigenvalues and orthogonal to the row vectors of M_G . In order to determine M_D with higher row number r , the resultant M_D with lower r is used with M_G to construct $Y(1)$. Using the IMSL routine ZXMWD for the global minimum with constraints one can determine these linear combination coefficients and consequently M_D .

Utilizing the concept of the minimal A -invariant subspace over a given subspace we developed a direct approach to determine the approximate lumping matrices. In the examples of the present paper, when \hat{n} is close to the dimension of the smallest simultaneously A_k -invariant subspace over \mathcal{H}_G , the resultant lumping matrices are the same as those with the global minimum error given by the first optimization approach. When the \hat{n} is low, the resultant lumping matrices are still close to the global minimum solutions given by the first approach. Therefore, one can employ the first optimization method to directly determine the elements of M instead of those of P and constrain the region of the unknown parameters around the solution given by the second direct approach.

Two examples used in our previous paper were employed to illustrate these new approaches. The results show that these new approaches are simpler and have higher accuracy than the method given in our previous paper. However, the approach presented

in our previous paper is general and can be applied in other cases, not only for the lumping schemes validated in the whole composition Y_n -space. The resultant M_D by the present paper might be used as an initial value for the matrix equations to determine the lumping schemes validated in any given region.

Acknowledgements—The authors acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

NOTATION

<i>Scalars</i>	
$a_k(y)$	k th coefficient of the decomposition of $J^T(y)$
C_i	i th species of a reaction system
k	integer
l	integer
m	integer
M_{ij}	(i, j) -entry of matrix M
\mathcal{H}	corresponding subspace of M
\mathcal{H}_G	corresponding subspace of M_G
\mathcal{V}	orthogonal direct complement of \mathcal{H} in n -dimensional space
n	dimension of vector y
\hat{n}	dimension of vector \hat{y}
P_{ij}	(i, j) -entry of matrix P
r	row number of M_D
\mathcal{R}^n	n -dimensional real space
s	integer
s_k	rank of A_k
t	time
Y_n	n -dimensional composition space
y_k	k th element of vector y

Z	total error defined as $\text{tr} \sum_{k=1}^m M A_k^T (I_n - M^T M) A_k M^T$	z	defined as $\begin{pmatrix} Y^T \\ M \end{pmatrix} y$
$Z(y)$	defined as $\text{tr} [E^T(y) E(y)]$	<i>Greek letters</i>	
Z_n	n -dimensional lumped species composition space	λ_i	i th eigenvalue of matrix $Y(1)$ or $Y(2)$
		Ω	desired region of the composition space

Vectors and matrices

Capital letters represent matrices; bold-face lower case letters represent vectors.

A	constant matrix
A_k	basis matrix of $J^T(y)$
B	constant matrix
$E(y)$	error matrix defined as $(I_n - M^T \bar{M}^T) J^T(y) M^T$
$f(y)$	n -dimensional function vector
$\hat{f}(\hat{y})$	\hat{n} -dimensional function vector
I	identity matrix
$J(y)$	Jacobian matrix of $f(y)$
\bar{M}	lumping matrix
M_D	determined submatrix of M
M_G	given submatrix of M
\bar{M}	generalized inverse of M satisfying $M\bar{M} = I_{\hat{n}}$
P	coefficient matrix
$Q_{(k)}^T$	matrix representation of $\text{Im}(A_k M^T)$ with orthonormal columns
$Q(G)_{(k)}^T$	matrix representation of $\text{Im}(A_k M_G^T)$ with orthonormal columns
$Q(D)_{(k)}^T$	matrix representation of $\text{Im}(A_k M_D^T)$ with orthonormal columns
$Q(G)_{(ki)}^T$	matrix representation of $\text{Im}[M_G(A_k^T)^i]^T$ with orthonormal columns
$Q(y)$	$\hat{n} \times \hat{n}$ function matrix
$R(1)$	eigenvector matrix of $Y(1)$
$R(2)$	eigenvector matrix of $Y(2)$
X	matrix representation of \mathcal{V} or submatrix of $R(1)$ and $R(2)$
y	n -dimensional variable vector
\hat{y}	\hat{n} -dimensional variable vector
$Y(1)$	symmetric matrix
$Y(2)$	symmetric matrix

Symbols

	any property related to the lumped system
0	null matrix

REFERENCES

- Aris, R., 1989, Reactions in continuous mixtures. *A.I.Ch.E. J.* **35**, 539–548.
- Astarita, G., 1989, Lumping nonlinear kinetics: apparent overall order of reaction. *A.I.Ch.E. J.* **35**, 529–532.
- Astarita, G. and Ocone, R., 1988, Lumping nonlinear kinetics. *A.I.Ch.E. J.* **34**, 1299–1309.
- Bellman, R., 1970, *Introduction to Matrix Analysis*. McGraw-Hill, New York.
- Ben-Israel, A. and Greville, T. N. E., 1974, *Generalized Inverse: Theory and Applications*. John Wiley, New York.
- Chou, M. Y. and Ho, T. C., 1988, Continuum theory for lumping nonlinear reaction mixtures. *A.I.Ch.E. J.* **34**, 1519–1527.
- Chou, M. Y. and Ho, T. C., 1989, Lumping coupled nonlinear reactions in continuous mixtures. *A.I.Ch.E. J.* **35**, 533–538.
- Coxson, P. G. and Bischoff, K. B., 1987a, Lumping strategy 1. Introduction techniques and applications of cluster analysis. *Ind. Engng Chem. Res.* **26**, 1239–1248.
- Coxson, P. G. and Bischoff, K. B., 1987b, Lumping strategy 2. A system theoretic approach. *Ind. Engng Chem. Res.* **26**, 2151–2157.
- Gohberg, I., Lancaster, P. and Rodman, L., 1986, *Invariant Subspaces of Matrices with Applications*. John Wiley, New York.
- Ho, T. C. and Aris, R., 1987, On apparent second-order kinetics. *A.I.Ch.E. J.* **33**, 1050–1051.
- Lang, S., 1986, *Introduction to Linear Algebra*, 2nd Edition. Springer, New York.
- Li, G. and Rabitz, H., 1989, A general analysis of exact lumping in chemical kinetics. *Chem. Engng Sci.* **44**, 1413–1430.
- Li, G. and Rabitz, H., 1990, A general analysis of approximate lumping in chemical kinetics. *Chem. Engng Sci.* **45**, 977–1002.

Appendix ii

8. A General Analysis of Exact Lumping in Chemical Kinetics, G. Li and H. Rabitz, Chem. Eng. Sci., 44, 1413 (1989).

A GENERAL ANALYSIS OF EXACT LUMPING IN CHEMICAL KINETICS

GENYUAN LI and HERSCHEL RABITZ

Department of Chemistry, Princeton University, Princeton, NJ 08540, U.S.A.

(Received 29 September 1987; accepted 12 September 1988)

Abstract—A general analysis of exact lumping is presented. This analysis can be applied to any reaction system with n species described by a set of first order ordinary differential equations $dy/dt = f(y)$, where y is an n -dimensional vector; $f(y)$ is an arbitrary n -dimensional function vector. Here we consider lumping by means of an $n \times n$ constant matrix M with rank n ($n < n$). It is found that a reaction system is exactly lumpable if and only if there exist nontrivial fixed invariant subspaces \mathcal{M} of the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$, no matter what value y takes, and the corresponding eigenvalues are the same for $J^T(y)$ and $J^T(My)$. Here the rows of M are the basis vectors of \mathcal{M} and \bar{M} is any generalized inverse of M satisfying $M\bar{M} = I_n$ with I_n being the n -identity matrix. The fixed invariant subspaces of $J^T(y)$ can be obtained either from the simultaneously invariant subspaces of all A_k , where the A_k 's form the basis of the decomposition of $J^T(y)$, or by determining the fixed $\text{Ker} \{ \Pi_i (J^T(y) - \lambda_i I_n)^{r_i} \Pi_j [(\sigma_j^2 + \tau_j^2) I_n - 2\sigma_j J^T(y) + (J^T(y))^2]^{r_j} \}$, where $\lambda_i, \sigma_j \pm i\tau_j$ are the real and nonreal eigenvalues of $J^T(y)$ and λ_i, σ_j and τ_j are usually functions of y ; r_i, r_j are nonnegative integers. The kinetic equations of the lumped system can be described as $d\bar{y}/dt = M\bar{f}(M\bar{y})$. This method is illustrated by some simple examples.

1. INTRODUCTION

A problem which frequently arises in the study of chemical kinetics is the high dimensionality and high degree of coupling of the reaction system. For example, in many realistic chemical processes, particularly those related to petrochemistry, industrial processes, combustion phenomena and atmospheric chemistry, the number of reacting species can often exceed 10^2 – 10^3 . It is impractical to incorporate the kinetic equations for each species. Consequently, lumping, by which several species are treated as a single component, is a necessity. Thus one desires to reduce the reaction mixture to a small number of lumps in the kinetic study for practical purposes. It is just as important to know how to systematically break down a model as it is to have the ability to build it up.

For different reaction systems the suitable ways of lumping will likely be different. Even for a given system, there could be many lumped models, depending on the objectives. However, one is not able to lump a system arbitrarily, because it is not always possible to find a model or a set of differential equations describing the behavior of the lumped species. For lack of the practical guidance, researchers have often spent many years trying to find adequate lumping schemes by trial and error. The modelling of catalytic cracking for petroleum (Jacob *et al.*, 1976) is a typical example. Confounding this approach is the fact that the true lumped "species" may actually be a combination or function of the original physical species.

Prior research clearly suggests the need for a rigorous study of lumping which can give useful guidelines for choosing lumps. Wei and Kuo (1965) were the first to give a lumping analysis of unimolecular reaction systems and their work was extended by Ozawa (1973) and Bailey (1972, 1975). One of the authors (Li,

1984) presented a lumping analysis for uni- and/or bimolecular reaction systems. Such research has been largely confined to uni- and/or bimolecular reaction systems with the focus on establishing the necessary and sufficient conditions for "exact lumping". These analyses have shown that exact lumping by a network of uni- and/or bimolecular reactions is feasible only under a very restrictive set of conditions. Studies of the pitfalls and magnitude of errors in the use of empirical rate expressions for lumping many independent single or consecutive reactions were presented by Luss and Hutchinson (1971), Luss (1975), Golikeri and Luss (1972, 1974) and Hutchinson and Luss (1970). Unfortunately until now lumping theory was not sufficiently developed to give useful guidelines as to which lumps to choose for many problems. There are still at least two important problems within exact lumping, which have not been solved yet.

- (1) There is no known *a priori* way to determine the lumping scheme.
- (2) The kinetic equations can have higher order nonlinearities than quadratic.

For instance, the second situation can arise in the presence of termolecular reactions. In addition, non-isothermal processes or the use of empirical rate laws can lead to highly nonlinear kinetic equations. Therefore, a general lumping analysis capable of treating arbitrary physical nonlinearities is necessary.

Considering this situation, a general analysis of exact lumping is presented in this paper. It can be used for any reaction system and the previously studied lumping analyses of uni- and/or bimolecular reaction systems are special cases of this analysis. In addition, this analysis can also be applied in other problems described by a set of first order ordinary differential

equations, such as problems arising in classical molecular dynamics, chemical engineering and control theory.

Section 2 of this paper presents the conditions under which a reaction system is exactly lumpable and the corresponding kinetic equations of the lumped system. In Section 3, the methods to determine the fixed invariant subspaces of the transpose of the Jacobian matrix of the kinetic equations are derived. Section 4 provides some simple examples to which the general lumping method is applied. Section 5 presents a discussion of the results.

2. CONDITIONS UNDER WHICH A REACTION SYSTEM IS EXACTLY LUMPABLE

Suppose the kinetics of an n -component reaction system can be described by

$$dy/dt = f(y), \quad (1)$$

where y is an n -composition vector; $f(y)$ is an arbitrary n -function vector, which does not contain t explicitly.

For practical purposes, here we only consider a special class of lumping by means of an $\hat{n} \times n$ real constant matrix M with rank \hat{n} ($\hat{n} < n$). If a system can be exactly lumped by the matrix M , it means that for

$$\hat{y} = My \quad (2)$$

we can find an \hat{n} -function vector $\hat{f}(\hat{y})$ such that

$$d\hat{y}/dt = \hat{f}(\hat{y}). \quad (3)$$

If y_i is not lumped, row i of M is the unit vector $e_i^T = (0 \dots 0 1 0 \dots 0)$, and $\hat{y}_i = y_i$. In this case, since the lumping is exact, the solutions for y_i and \hat{y}_i by eqs (1) and (3) are the same. However, eq. (3) is simpler.

Not every system is exactly lumpable. Therefore, we need to determine the necessary and sufficient conditions for the existence of exact lumping. We also desire that these conditions be constructive in order to determine the lumping matrices. From eqs (1) and (2) we have

$$d\hat{y}/dt = M dy/dt = Mf(y), \quad (4)$$

and upon comparing eqs (3) and (4) we have

$$\hat{f}(\hat{y}) = Mf(y). \quad (5)$$

As the rank of M is \hat{n} , there must exist generalized inverses (Isral, 1974) \bar{M} of matrix M satisfying

$$M\bar{M} = I_{\hat{n}}, \quad (6)$$

where $I_{\hat{n}}$ is the \hat{n} -identity matrix. Substituting eq. (2) into eq. (5) yields

$$\hat{f}(My) = Mf(y), \quad (7)$$

and this is an identity for any y . Therefore, letting

$$y = M\hat{y}, \quad (8)$$

we have

$$\hat{f}(MM\hat{y}) = Mf(M\hat{y}),$$

$$\hat{f}(\hat{y}) = Mf(M\hat{y}) \quad (9)$$

Comparing eqs (5) and (9), we obtain the necessary condition for the existence of exact lumping

$$Mf(y) = Mf(M\hat{y}),$$

$$Mf(y) = Mf(M\hat{y}). \quad (10)$$

Equation (10) is also sufficient for the existence of exact lumping. Indeed, if we choose

$$\hat{f}(\hat{y}) = Mf(M\hat{y}),$$

then the behavior of the lumped species can be described by

$$d\hat{y}/dt = Mf(M\hat{y}), \quad (11)$$

and according to eq. (10) the lumped system satisfies eq. (4). Then we have

$$d\hat{y}/dt = M dy/dt,$$

$$d(\hat{y} - My)/dt = 0,$$

$$\hat{y} - My = c,$$

where c is an arbitrary constant vector. Choosing $c = 0$ gives

$$\hat{y} = My.$$

Equation (10) does not place any restriction on \bar{M} except that $M\bar{M} = I_{\hat{n}}$. This latter point is important in that the nonunique nature of \bar{M} does not effect the form of the lumped equations (physical model) in the exact case. It means that \bar{M} in eq. (11) is any one of the generalized inverses satisfying $M\bar{M} = I_{\hat{n}}$. This can be easily demonstrated as follows.

Considering once again that eq. (10) is an identity for all y , let y take the following value

$$\bar{M}'My,$$

where \bar{M}' is another generalized inverse of M . We get

$$\begin{aligned} Mf(\bar{M}'My) &= Mf(M\bar{M}'My), \\ &= Mf(M\bar{M}y) \end{aligned}$$

or

$$Mf(\bar{M}'y) = Mf(\bar{M}y). \quad (12)$$

This shows that different generalized inverses of M give the same lumped model.

We cannot directly apply eq. (10) to examine whether a system is exactly lumpable or not, because we do not know M in advance. In order to obtain further insight into exact lumping, we differentiate both sides of eq. (10) with respect to y to produce

$$MJ(y) = MJ(M\hat{y})\bar{M}. \quad (13)$$

Since the rank of M is \hat{n} , it has a nontrivial null space \mathcal{N} with dimension $n - \hat{n}$. We can verify that \mathcal{N} is invariant under $J(y)$, no matter what value y takes. Indeed, for every $x \in \mathcal{N}$, we have

$$MJ(y)x = MJ(M\hat{y})\bar{M}Mx = 0 \quad (14)$$

This implies that $J(y)x \in \mathcal{N}$ for any value of y , so \mathcal{N} is $J(y)$ -invariant.

Suppose \mathcal{V} is represented as

$$\mathcal{V} = \text{Span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-n_0}\}, \quad (15)$$

where \mathbf{x}_i 's are the basis of \mathcal{V} . Let vectors \mathbf{x}_i compose the columns of matrix X , then

$$MX = 0, \quad (16)$$

and

$$MJ(\mathbf{y})X = 0. \quad (17)$$

Note that if \mathcal{V} is $J(\mathbf{y})$ -invariant, then \mathcal{V}^\perp is $J^T(\mathbf{y})$ -invariant (Gohberg *et al.*, 1986). Let $\mathcal{H} = \mathcal{V}^\perp$. Considering eq. (16), it is obvious that \mathcal{H} is spanned by the row vectors of M .

$$\mathcal{H} = \text{Span}\{\mathbf{m}_{(1)}, \mathbf{m}_{(2)}, \dots, \mathbf{m}_{(n_0)}\}, \quad (18)$$

where $\mathbf{m}_{(i)}$ is the transpose of row i of M .

In conclusion, a system described as eq. (1) can be exactly lumped by an $n \times n$ real constant matrix M , only if the nullspace \mathcal{V} of M is $J(\mathbf{y})$ -invariant or the subspace \mathcal{H} spanned by the row vectors of M is $J^T(\mathbf{y})$ -invariant, no matter what value \mathbf{y} takes. We call \mathcal{H} and \mathcal{V} the fixed (i.e. \mathbf{y} independent) invariant subspaces of variable matrices. Since \mathcal{H} and \mathcal{V} are orthogonal complements, each one can be obtained when the other has been determined. In order to determine M directly we mainly consider \mathcal{H} in the following analysis. However, the existence of the $J(\mathbf{y})$ - or $J^T(\mathbf{y})$ -fixed invariant subspaces is only a necessary condition, i.e. not every M corresponding to \mathcal{H} can be used as a lumping matrix. We need to find the condition under which \mathcal{H} can supply a lumping matrix, and this result is established below.

It is well known that a subspace $\mathcal{H} = \text{Span}\{\mathbf{m}_{(1)}, \mathbf{m}_{(2)}, \dots, \mathbf{m}_{(n_0)}\}$ is $J^T(\mathbf{y})$ -invariant if and only if $J^T(\mathbf{y})\mathbf{m}_{(k)} \in \mathcal{H}$, i.e. the image of $\mathbf{m}_{(k)}$ upon mapping by matrix $J^T(\mathbf{y})$ is a certain linear combination of all $\mathbf{m}_{(i)}$:

$$J^T(\mathbf{y})\mathbf{m}_{(k)} = \sum_{i=1}^{n_0} q_{ki}(\mathbf{y})\mathbf{m}_{(i)}, \quad (19)$$

where $q_{ki}(\mathbf{y})$'s are the linear combination coefficients, which are usually functions of \mathbf{y} . Considering all $\mathbf{m}_{(i)}$ gives

$$J^T(\mathbf{y})M^T = M^T Q^T(\mathbf{y}).$$

Transposing it yields

$$MJ(\mathbf{y}) = Q(\mathbf{y})M, \quad (20)$$

where $Q(\mathbf{y})$ is an $n \times n$ matrix with $q_{ij}(\mathbf{y})$ as its (i, j) -entry. Since \mathcal{H} is invariant under $J^T(\mathbf{y})$ for any value of \mathbf{y} , therefore we also have

$$MJ(\bar{M}\mathbf{y}) = Q(\bar{M}\mathbf{y})M. \quad (21)$$

Using this relation, we can deduce the sufficient condition for exact lumping. Note that

$$\begin{aligned} MJ(\bar{M}\mathbf{y})\bar{M}M &= Q(\bar{M}\mathbf{y})\bar{M}MM \\ &= Q(\bar{M}\mathbf{y})M \end{aligned} \quad (22)$$

Comparing eqs (21) and (22) yields

$$MJ(\bar{M}\mathbf{y}) = MJ(\bar{M}\mathbf{y})\bar{M}M \quad (23)$$

Substituting eq. (23) into eq. (13) and rearranging it, we obtain another necessary condition for the existence of exact lumping:

$$M[J(\mathbf{y}) - J(\bar{M}\mathbf{y})] = 0. \quad (24)$$

It is easy to prove that eq. (24) is also sufficient for exact lumping, if, in addition, \mathcal{H} is $J^T(\mathbf{y})$ -invariant for any \mathbf{y} . Indeed, when \mathcal{H} is $J^T(\mathbf{y})$ -invariant for any \mathbf{y} , eq. (23) holds. Consequently, eq. (23) and (24) give eq. (13). We now write eq. (13) in the explicit form

$$\begin{aligned} M \begin{pmatrix} \frac{\partial f_1(\mathbf{y})}{\partial y_1} & \frac{\partial f_1(\mathbf{y})}{\partial y_2} & \dots & \frac{\partial f_1(\mathbf{y})}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{y})}{\partial y_1} & \frac{\partial f_n(\mathbf{y})}{\partial y_2} & \dots & \frac{\partial f_n(\mathbf{y})}{\partial y_n} \end{pmatrix} \\ = M \begin{pmatrix} \frac{\partial f_1(\mathbf{z})}{\partial z_1} & \frac{\partial f_1(\mathbf{z})}{\partial z_2} & \dots & \frac{\partial f_1(\mathbf{z})}{\partial z_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{z})}{\partial z_1} & \frac{\partial f_n(\mathbf{z})}{\partial z_2} & \dots & \frac{\partial f_n(\mathbf{z})}{\partial z_n} \end{pmatrix} \bar{M}M, \end{aligned} \quad (25)$$

where $\mathbf{z} = \bar{M}\mathbf{y}$, z_i is the i th element of \mathbf{z} . Multiplying both sides of eq. (25) from the right by $d\mathbf{y}$ and integrating give

$$M \begin{pmatrix} \int \sum_{i=1}^n \frac{\partial f_1(\mathbf{y})}{\partial y_i} dy_i \\ \vdots \\ \int \sum_{i=1}^n \frac{\partial f_n(\mathbf{y})}{\partial y_i} dy_i \end{pmatrix} = M \begin{pmatrix} \int \sum_{i=1}^n \frac{\partial f_1(\mathbf{z})}{\partial z_i} dz_i \\ \vdots \\ \int \sum_{i=1}^n \frac{\partial f_n(\mathbf{z})}{\partial z_i} dz_i \end{pmatrix}. \quad (26)$$

Since the total differential of $f_j(\mathbf{y})$ is

$$df_j(\mathbf{y}) = \sum_{i=1}^n \frac{\partial f_j(\mathbf{y})}{\partial y_i} dy_i, \quad (27)$$

eq. (26) implies that

$$M \begin{pmatrix} \int df_1(\mathbf{y}) \\ \vdots \\ \int df_n(\mathbf{y}) \end{pmatrix} = M \begin{pmatrix} \int df_1(\mathbf{z}) \\ \vdots \\ \int df_n(\mathbf{z}) \end{pmatrix}.$$

$$M\mathbf{f}(\mathbf{y}) - M\mathbf{f}(\mathbf{z}) = \mathbf{c},$$

$$M\mathbf{f}(\mathbf{y}) = M\mathbf{f}(\bar{M}\mathbf{y}) + \mathbf{c}, \quad (28)$$

where \mathbf{c} is n -dimensional arbitrary constant vector. If we choose $\mathbf{c} = \mathbf{0}$, we obtain eq. (10), which is the necessary and sufficient condition for exact lumping. In addition, it can also be shown that eq. (13) is a

necessary and sufficient condition for the existence of exact lumping.

This necessary and sufficient condition can be described in an alternative way. Substituting eqs (20) and (21) into eq. (24) yields

$$[Q(y) - Q(\bar{M}My)]M = 0. \quad (29)$$

Since M is a row-full rank matrix, we can always find \hat{n} columns from it to construct a nonsingular $\hat{n} \times \hat{n}$ matrix M_{sq} such that

$$[Q(y) - Q(\bar{M}My)]M_{sq} = 0. \quad (30)$$

Transposing this equation gives

$$M_{sq}^T [Q^T(y) - Q^T(\bar{M}My)] = 0. \quad (31)$$

Considering that M_{sq}^T is nonsingular, its null space is only $\{0\}$. Therefore, we have

$$Q^T(y) - Q^T(\bar{M}My) = 0, \quad (32)$$

or

$$Q^T(y) = Q^T(\bar{M}My). \quad (33)$$

If we consider y symbolically, $Q^T(y)$ can be treated in the same way as that of a constant matrix. It is well known that there is a Jordan canonical form related to $Q^T(y)$ (Appendix A):

$$Q^T(y) = S(y)J_{or}[\lambda(y)]S^{-1}(y), \quad (34)$$

where $S(y)$ is an invertible matrix, i.e. the determinant of $S(y)$ is not identically equal to zero for all y and $J_{or}[\lambda(y)]$ is the Jordan matrix (Gohberg *et al.*, 1986).

After transposing and considering eq. (34), eq. (20) becomes

$$J^T(y)M^T = M^T S(y)J_{or}[\lambda(y)]S^{-1}(y). \quad (35)$$

Multiplying both sides of the above equation from the right by $S(y)$ yields

$$J^T(y)M^T S(y) = M^T S(y)J_{or}[\lambda(y)], \quad (36)$$

$$J^T(y)M'^T = M'^T J_{or}[\lambda(y)], \quad (37)$$

where

$$M'^T = M^T S(y). \quad (38)$$

M' has rank \hat{n} , because $S(y)$ is nonsingular. Since the rows of M' are linear combinations of those of M , then the row vectors of M' are just another basis of \mathcal{H} . The elements of $\lambda(y)$ are the subset of the eigenvalues of $J^T(y)$ corresponding to \mathcal{H} .

A companion formula to eq. (37) can be obtained by considering

$$Q^T(y) = Q^T(\bar{M}My),$$

in eq. (21) to give

$$J^T(\bar{M}My)M'^T = M'^T J_{or}[\lambda(y)]. \quad (39)$$

Equations (37) and (39) imply that when a system is exactly lumpable, the eigenvalues of $J^T(\bar{M}My)$ corresponding to the fixed invariant subspace \mathcal{H} will be the same as those of $J^T(y)$. This is also sufficient for exact lumping, because eqs (37) and (39) will give

$$J^T(y)M'^T = J^T(\bar{M}My)M'^T. \quad (40)$$

Multiplying both sides of this equation from the right by $S^{-1}(y)$ yields eq. (24), which has been proved to be a necessary and sufficient condition for exact lumping. Therefore, the alternative description of the necessary and sufficient condition for exact lumping obtained by eq. (24) is the following: a system is exactly lumpable if and only if its $J^T(y)$ has nontrivial fixed invariant subspaces and the corresponding eigenvalues for $J^T(y)$ and $J^T(\bar{M}My)$ are the same.

When the corresponding eigenvalues of a fixed invariant subspace \mathcal{H} for $J^T(y)$ are not functions of y , i.e. they are constants, then it always holds that $J^T(y)$ and $J^T(\bar{M}My)$ have the same eigenvalues corresponding to \mathcal{H} . However, the presence of constant eigenvalues cannot guarantee the existence of exact lumping, because sometimes one cannot find a fixed invariant subspace of $J^T(y)$ related to these constant eigenvalues. It is easy to give an example of this. Consider the matrix

$$A(y) = \begin{pmatrix} y_1 + 2 & y_2 \\ y_1 & y_2 + 2 \end{pmatrix}.$$

The eigenvalues of $A(y)$ are 2 and $y_1 + y_2 + 2$. The corresponding eigenvectors are

$$\begin{pmatrix} -y_2 \\ y_1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

respectively. One can see that the constant eigenvalue 2 does not have a fixed eigenvector. In contrast, $y_1 + y_2 + 2$ does have a constant one.

As a special case, when a system is linear, $J^T(y)$ is a constant matrix. In this situation, the fixed invariant subspaces exist and they correspond to constant eigenvalues. Therefore, a linear system is always exactly lumpable and any $J^T(y)$ -invariant subspace will give a lumping matrix.

When $Q(y)$ is a constant matrix Q , it is interesting that the lumped system is linear, no matter if the original system is linear or not. In this case, eq. (34) becomes

$$Q^T = S J_{or}(\lambda) S^{-1}, \quad (41)$$

and all eigenvalues are constants, i.e. the fixed invariant subspace \mathcal{H} of $J^T(y)$ is related to constant eigenvalues. Equation (20) then becomes

$$MJ(y) = QM. \quad (42)$$

Multiplying both sides of eq. (42) by dy and integrating under an appropriate integration condition give

$$Mf(y) = QMy, \quad (43)$$

i.e.

$$d\hat{y} \cdot dt = Q\hat{y}, \quad (44)$$

which are linear differential equations.

In summary, for exact lumping, (i) we need to determine whether the fixed nontrivial invariant subspaces \mathcal{H} of $J^T(y)$ exist or not; (ii) if they do exist, then we need to examine whether they satisfy either eq. (10).

(13), (24) or the corresponding eigenvalues for $J^T(\mathbf{y})$ and $J^T(\bar{M}M\mathbf{y})$ are the same. When these two conditions are satisfied, the system described as eq. (1) is exactly lumpable by matrix M , whose rows are composed of the basis vectors of \mathcal{H} .

3. DETERMINATION OF THE FIXED $J^T(\mathbf{y})$ -INVARIANT SUBSPACES \mathcal{H}

In order to determine lumping matrices M we need first to determine the fixed $J^T(\mathbf{y})$ -invariant subspaces \mathcal{H} . There are two ways to determine them. Before discussing these approaches, we first consider the decomposition of $J^T(\mathbf{y})$, which will be important for implementing the determination of \mathcal{H} .

(A) Decomposition of $J^T(\mathbf{y})$

The Jacobian matrix can be considered as an n^2 -vector. Therefore, for any value of \mathbf{y} , $J^T(\mathbf{y})$ can be represented as a linear combination of m ($m \leq n^2$) constant matrices

$$J^T(\mathbf{y}) = \sum_{k=1}^m a_k(\mathbf{y}) A_k, \quad (45)$$

where $a_k(\mathbf{y})$ are parameters, which are functions of \mathbf{y} ; the A_k 's are constant matrices considered as a basis of $J^T(\mathbf{y})$. The problem is how to determine the basis A_k 's. There are several ways to achieve this task, and one is as follows. The variable $J^T(\mathbf{y})$ can be represented as

$$J^T(\mathbf{y}) = \sum_{i,j=1}^n j_{ij}(\mathbf{y}) E_{ij}, \quad (46)$$

where $j_{ij}(\mathbf{y})$ is the (i, j) -entry of $J^T(\mathbf{y})$; E_{ij} is the elementary matrix, which is defined as the $n \times n$ matrix having unity in the (i, j) th position and all other elements are zero (Graham, 1981). If $j_{pq}(\mathbf{y})$ is equal to $c j_{ij}(\mathbf{y})$, where c is a constant, we can combine these two terms as

$$a_k(\mathbf{y}) = j_{ij}(\mathbf{y}), \quad (47)$$

$$A_k = E_{ij} + c E_{pq}. \quad (48)$$

In this way one can combine as many terms as possible in eq. (46) to obtain eq. (45), where m is less than n^2 . The remainder of this section is concerned with the determination of \mathcal{H} .

(B) Approach 1 to determine \mathcal{H}

It is easy to demonstrate that the simultaneously invariant subspaces of all constant matrices A_k 's are $J^T(\mathbf{y})$ -invariant. To establish this point let \mathcal{H} represent a simultaneously invariant subspace for all A_k 's, i.e. for every $\mathbf{x} \in \mathcal{H}$, we have $A_k \mathbf{x} \in \mathcal{H}$ for all k . Using this relation, we obtain

$$J^T(\mathbf{y}) \mathbf{x} = \sum_{k=1}^m a_k(\mathbf{y}) A_k \mathbf{x} \in \mathcal{H}. \quad (49)$$

Equation (49) shows that \mathcal{H} is invariant under $J^T(\mathbf{y})$.

If eq. (45) satisfies the restriction that we can choose an appropriate value \mathbf{y}_i of \mathbf{y} such that all $a_k(\mathbf{y}_i)$'s vanish

except $a_i(\mathbf{y}_i)$, i.e.

$$J^T(\mathbf{y}_i) = a_i(\mathbf{y}_i) A_i, \quad (i = 1, 2, \dots, m) \quad (50)$$

then the fixed $J^T(\mathbf{y})$ -invariant subspaces are also simultaneously invariant for all A_k 's. Indeed, if \mathcal{H} is $J^T(\mathbf{y})$ -invariant for all values of \mathbf{y} , it must be invariant under $J^T(\mathbf{y}_i)$, i.e. for every $\mathbf{x} \in \mathcal{H}$ we also have $J^T(\mathbf{y}_i) \mathbf{x} \in \mathcal{H}$. Since $a_i(\mathbf{y}_i)$ is not equal to zero, then

$$A_i = J^T(\mathbf{y}_i)/a_i(\mathbf{y}_i). \quad (51)$$

For every $\mathbf{x} \in \mathcal{H}$, we have

$$A_i \mathbf{x} = J^T(\mathbf{y}_i) \mathbf{x} / a_i(\mathbf{y}_i) \in \mathcal{H}. \quad (i = 1, 2, \dots, m) \quad (52)$$

This result shows that \mathcal{H} is simultaneously invariant for all A_k 's. Thus we can determine the invariant subspaces of $J^T(\mathbf{y})$ by only determining the simultaneously invariant subspaces of all A_k 's. We should emphasize that this restriction is sufficient, but not necessary for $J^T(\mathbf{y})$ -invariant subspaces to be simultaneously invariant under all A_k .

When a reaction system is uni- and/or bimolecular, the elements of $J^T(\mathbf{y})$ are only linear functions of the y_k 's. In this case, eq. (45) will have a simple form, i.e. $a_k(\mathbf{y})$ is either constant or y_k ,

$$J^T(\mathbf{y}) = A_0 + \sum_{k=1}^m y_k A_k, \quad (53)$$

where m is equal to or less than n , and A_0 can be the null matrix. It is easy to prove that the fixed $J^T(\mathbf{y})$ -invariant subspaces are simultaneously A_0 - and all A_k -invariant. Suppose \mathcal{H} is a fixed $J^T(\mathbf{y})$ -invariant subspace for any value of \mathbf{y} . Therefore, \mathcal{H} must be invariant to $J^T(\mathbf{0})$. Equation (53) gives

$$J^T(\mathbf{0}) = A_0.$$

For every $\mathbf{x} \in \mathcal{H}$, we have

$$A_0 \mathbf{x} = J^T(\mathbf{0}) \mathbf{x} \in \mathcal{H}, \quad (54)$$

which implies that \mathcal{H} is A_0 -invariant. Similarly, \mathcal{H} is $J^T(\mathbf{e}_i)$ -invariant. Equation (53) gives

$$J^T(\mathbf{e}_i) = A_0 + A_i.$$

$$A_i = J^T(\mathbf{e}_i) - A_0.$$

For every $\mathbf{x} \in \mathcal{H}$, we have

$$A_i \mathbf{x} = J^T(\mathbf{e}_i) \mathbf{x} - A_0 \mathbf{x} \in \mathcal{H}. \quad (i = 1, 2, \dots, m) \quad (55)$$

One can see that \mathcal{H} is simultaneously all A_k -invariant ($k = 0, 1, \dots, m$). Therefore, we can determine the fixed invariant subspaces of $J^T(\mathbf{y})$ by determining the simultaneously invariant ones of all A_k 's.

Suppose \mathcal{H} is a subspace, which is simultaneously invariant for all A_k . It is easy to demonstrate that \mathcal{H} is also invariant under $\sum_{k=0}^m A_k$ and $\prod_{k=0}^m A_k$. For a transformation A , we denote by $\text{Inv}(A)$ the set of all A -invariant subspaces, including the null subspace $\{0\}$ and the n -dimensional space \mathcal{H}^n . Then we have the conclusion that all simultaneously invariant subspaces for all A_k 's are contained in $\text{Inv}(\sum_{k=0}^m A_k)$ and $\text{Inv}(\prod_{k=0}^m A_k)$. $\sum_{k=0}^m A_k$ and $\prod_{k=0}^m A_k$ are constant matrices, and $\text{Inv}(\sum_{k=0}^m A_k)$ or $\text{Inv}(\prod_{k=0}^m A_k)$ can be

easily determined through their Jordan canonical form. For any constant matrix A , there is a biggest A -invariant subspace called the root subspace corresponding to each eigenvalue of A . Using the Jordan canonical form all A -invariant subspaces contained in each root subspace can be readily determined, and all the sums of the A -invariant subspaces in different root subspaces compose the full set $\text{Inv}(A)$ (Appendix B). The invariant subspaces of $J^T(y)$ can be obtained by examining which subspaces in $\text{Inv}(\sum_{k=0}^m A_k)$ or $\text{Inv}(\prod_{k=0}^m A_k)$ are simultaneously invariant for all A_k 's. One can achieve this task by examining whether the image vectors of the basis vectors of a subspace in $\text{Inv}(\sum_{k=0}^m A_k)$ or $\text{Inv}(\prod_{k=0}^m A_k)$ upon mapping by A_k are still in the same subspace, i.e. any image vector can be represented as a certain linear combination of the basis vectors of this subspace.

(C) Approach II to determine \mathcal{H}

There is another way to determine the fixed $J^T(y)$ -invariant subspaces. Let $\text{Ker } A^r$ represent the null subspace of A^r . We know that the $\text{Ker}(A - \lambda_i I_n)^r$ and $\text{Ker}[\prod_{i=1}^k (A - \lambda_i I_n)^{r_i}]$ of a linear transformation A are A -invariant, where r, r_i are positive integers and

$$\text{Ker}(A - \lambda_i I_n)^r \subset \text{Ker}(A - \lambda_i I_n)^{r+1}. \quad (56)$$

Since the dimension of $\text{Ker}(A - \lambda_i I_n)^r, r = 1, 2, \dots$ are bounded above by n , there exists a minimal integer $p_i \geq 1$ such that

$$\text{Ker}(A - \lambda_i I_n)^r \subset \text{Ker}(A - \lambda_i I_n)^{p_i} \quad (57)$$

for all positive integers r . $\text{Ker}(A - \lambda_i I_n)^{p_i}$ is called the root subspace of A corresponding to λ_i and is denoted by $\mathcal{H}_{\lambda_i}(A)$.

Therefore, solving the equation

$$(A - \lambda_i I_n)^r \mathbf{x} = \mathbf{0} \quad (r_i = 1, 2, \dots, p_i) \quad (58)$$

for each eigenvalue will give A -invariant subspaces with different dimensions. In addition, we also have

$$\text{Ker} \left[\prod_{i=1}^k (A - \lambda_i I_n)^{r_i} \right] \subset \text{Ker} \left[\prod_{i=1}^{k+1} (A - \lambda_i I_n)^{r_i} \right]. \quad (59)$$

We need to solve the following equations to obtain all A -invariant subspaces with different dimensions:

$$\prod_{i=1}^k (A - \lambda_i I_n)^{r_i} \mathbf{x} = \mathbf{0}, \quad (k = 1, 2, \dots, t; \quad r_i = 0, 1, \dots, p_i) \quad (60)$$

where t is the number of the distinct eigenvalues of A . Here we define

$$(A - \lambda_i I_n)^0 = I_n. \quad (61)$$

Sometimes A has nonreal eigenvalues $\sigma_1 \pm i\tau_1, \dots, \sigma_s \pm i\tau_s$. For our purposes here, we aim only to obtain real lumping matrices (this restriction may be removed, if desired). Therefore, we need to determine the real null subspaces for nonreal eigenvalues. In order to do so, we consider the null subspace $\text{Ker}[(\sigma_i^2 + \tau_i^2)I_n - 2\sigma_i A + A^2]^{q_i}$ for $\sigma_i \pm i\tau_i$, and the

corresponding root subspace is defined as

$$\mathcal{H}_{\sigma_i \pm i\tau_i}(A) = \text{Ker}[(\sigma_i^2 + \tau_i^2)I_n - 2\sigma_i A + A^2]^{q_i}. \quad (62)$$

The determination of A -invariant subspaces corresponding to nonreal eigenvalues is similar to that of real eigenvalues.

This approach can be applied to determine the fixed $J^T(y)$ -invariant subspaces. Let $\lambda_1, \dots, \lambda_t, \sigma_1 \pm i\tau_1, \dots, \sigma_s \pm i\tau_s$ be all distinct real and nonreal eigenvalues of $J^T(y)$. Here $\lambda_i, \sigma_i, \tau_i$ are usually functions of y . We solve the following equations to find the constant vector solutions \mathbf{x} 's, if they exist.

$$\left\{ \prod_{i=1}^k [J^T(y) - \lambda_i I_n]^{r_i} \prod_{j=1}^{k'} [(\sigma_j^2 + \tau_j^2)I_n - 2\sigma_j J^T(y) + (J^T(y))^2]^{r_j} \right\} \mathbf{x} = \mathbf{0}. \quad (63)$$

$$(k = 1, 2, \dots, t; \quad k' = 1, 2, \dots, s; \quad r_i = 0, 1, \dots, p_i;$$

$$r_j = 0, 1, \dots, q_j)$$

The subspaces spanned by the linearly independent constant solution \mathbf{x} 's of eq. (63) give the fixed $J^T(y)$ -invariant subspaces with different dimensions.

Notice the difference between eq. (63) and the preceding discussion for a constant matrix A . In eq. (63) $J^T(y)$ is a variable matrix and \mathbf{x} are constrained to be constant vectors. Therefore, in this situation we can not apply the concept of root subspace directly. The largest values of r_i and r_j perhaps may not be equal to p_i and q_j , respectively.

A difficulty can arise, when eq. (63) contains all distinct eigenvalues of $J^T(y)$. The product of the matrices on the left side of eq. (63) can be related to the minimal polynomial of $J^T(y)$, and then it becomes the null matrix (Appendix C). In this case, any vector is a solution of eq. (63) and we can do nothing with it. However, notice that when this situation arises, the constant solutions correspond to the fixed $J^T(y)$ -invariant subspaces with the highest dimensions. We know that the orthogonal complementary subspaces of $J(y)$ -invariant subspaces are $J^T(y)$ -invariant. The sum of the dimensions for a subspace and its complementary one is n . Therefore, we can first determine the fixed $J(y)$ -invariant subspaces with the lowest dimensions by the same way. Then their orthogonal complementary subspaces give the fixed $J^T(y)$ -invariant ones with the highest dimensions.

The Approaches I and II outlined above to determine the fixed $J^T(y)$ -invariant subspaces will be illustrated by uni- and/or bimolecular reactions below.

4. APPLICATION TO UNI- AND/OR BIMOLECULAR REACTION SYSTEMS

As examples of the application of the analysis above, we choose uni- and/or bimolecular reaction systems. In this case the transpose of the Jacobian matrix can be described as

$$J^T(y) = A_0 + \sum_{k=1}^n v_k A_k. \quad (64)$$

For a unimolecular reaction system, the kinetic equations are

$$dy/dt = Ky, \quad (65)$$

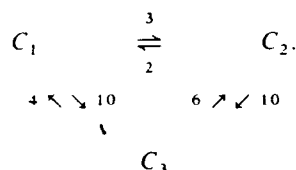
where K is the rate constant matrix. The Jacobian matrix for the unimolecular reaction system is just K , and then

$$J^T(y) = K^T. \quad (66)$$

For realistic chemical kinetics all eigenvalues of K (or K^T) are nonpositive real numbers (Wei and Prater, 1963).

Example 1

A unimolecular reaction system with 3 species (Wei and Kuo, 1969) is described as follows:



where C_1 , C_2 and C_3 represent the three species; all numbers are unitless rate constants. Let y_i represent the concentration of species C_i . Then the corresponding kinetic equations can be described as eq. (65) and

$$J^T(y) = K^T = \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix}^T. \quad (67)$$

The eigenvector matrix X and the eigenvalue matrix Λ of K^T are

$$X = \begin{pmatrix} 1 & 1 & 0.6 \\ 1 & 1 & -0.4 \\ 1 & -1 & 0 \end{pmatrix}, \quad (68)$$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -20 & 0 \\ 0 & 0 & -15 \end{pmatrix}. \quad (69)$$

From Section 2 we know that any linear system is exactly lumpable and any invariant subspace of $J^T(y)$ can be used to construct a lumping matrix. Then the only thing we need to do is determining all of the K^T -invariant subspaces, whose basis vectors compose the lumping matrices. Considering that the eigenvalues of K^T are distinct, any subspace spanned by a subset of its eigenvectors is invariant to it. For convenience let x_1 , x_2 and x_3 represent the 3 columns of X . Then $\text{Inv}(K^T)$ contains

$$\begin{aligned} &\text{Span}\{0\}, \text{Span}\{x_1\}, \text{Span}\{x_2\}, \text{Span}\{x_3\}, \\ &\text{Span}\{x_1, x_2\}, \text{Span}\{x_1, x_3\}, \text{Span}\{x_2, x_3\}, \\ &\mathcal{R}^3. \end{aligned}$$

The number of K^T -invariant subspaces is finite, but the number of the lumping matrices is infinite, because one can choose different bases to represent 2-dimensional invariant subspaces. For example, $\text{Span}\{x_1$,

$x_2\}$ can give the lumping matrix

$$M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{pmatrix}. \quad (70)$$

We can also obtain another lumping matrix by elementary row operations (Lang, 1986) on the two rows:

$$\bar{M} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (71)$$

The rows of the new lumping matrix are just another basis of the same invariant subspace.

In Section II we proved that the nonunique nature of \bar{M} does not effect the form of the lumped equations. For the \bar{M} given in eq. (71), for example, we can find an infinite number of \bar{M} satisfying $M\bar{M} = I_2$. We arbitrarily choose two:

$$\bar{M}_1 = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{pmatrix}, \quad \bar{M}_2 = \begin{pmatrix} 0.4 & 0 \\ 0.6 & 0 \\ 0 & 1 \end{pmatrix}. \quad (72)$$

It is easy to show that the kinetic equations for the lumped system are the same in spite of using different \bar{M} . According to eq. (11)

$$\hat{f}(\hat{y}) = Mf(\bar{M}\hat{y}),$$

and since

$$f(y) = Ky,$$

then

$$\hat{f}(\hat{y}) = MK\bar{M}\hat{y}. \quad (73)$$

For \bar{M}_1 we have

$$\begin{aligned} \hat{f}(\hat{y}) &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{pmatrix} \hat{y} \\ &= \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix} \hat{y}. \end{aligned}$$

Similarly for \bar{M}_2 we have

$$\begin{aligned} \hat{f}(\hat{y}) &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \begin{pmatrix} 0.4 & 0 \\ 0.6 & 0 \\ 0 & 1 \end{pmatrix} \hat{y} \\ &= \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix} \hat{y}. \end{aligned}$$

They give the same kinetic equations for the lumped system, whose reaction scheme can be described as

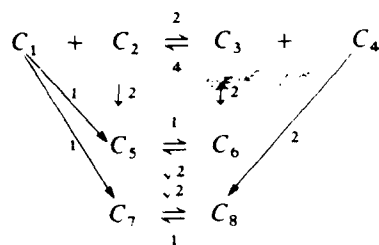


where $\hat{y} = (\hat{y}_1, \hat{y}_2)^T$ is the concentration vector of \hat{C}_1 , \hat{C}_2 and

$$\hat{K} = \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix}. \quad (75)$$

Example 2

A uni- and bimolecular reaction system with 8 species (Li, 1984) is illustrated as follows:



where the C_i 's are species; the numbers are unitless rate constants.

Letting y_i represent the concentration of C_i , it is easy to write out the kinetic equations and the transpose of the corresponding Jacobian matrix $J^T(y)$.

$$\begin{aligned}
 dy_1/dt &= -2y_1 - 2y_1y_2 + 4y_3y_4 \\
 dy_2/dt &= -2y_2 - 2y_1y_2 + 4y_3y_4 \\
 dy_3/dt &= -2y_3 - 4y_3y_4 + 2y_1y_2 \\
 dy_4/dt &= -2y_4 - 4y_3y_4 + 2y_1y_2 \\
 dy_5/dt &= -y_5 + y_1 + 2y_2 + \sqrt{2}y_6 \\
 dy_6/dt &= -\sqrt{2}y_6 + 2y_3 + y_5 \\
 dy_7/dt &= -\sqrt{2}y_7 + y_1 + y_8 \\
 dy_8/dt &= -y_8 + 2y_4 + \sqrt{2}y_7
 \end{aligned} \quad (76)$$

$$J^T(y) = \begin{pmatrix} -2(1+y_2) & -2y_2 & 2y_2 & 2y_2 & 1 & 0 & 1 & 0 \\ -2y_1 & -2(1+y_1) & 2y_1 & 2y_1 & 2 & 0 & 0 & 0 \\ 4y_4 & 4y_4 & -2(1+2y_4) & -4y_4 & 0 & 2 & 0 & 0 \\ 4y_3 & 4y_3 & -4y_3 & -2(1+2y_3) & 0 & 0 & 0 & 2 \\ & & & & -1 & 1 & 0 & 0 \\ & & & & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ & & & & 0 & 0 & 1 & -1 \end{pmatrix}$$

According to Section 2, in order to determine the lumping matrices we need first to establish all the fixed $J^T(y)$ -invariant subspaces. This task can be done by Approaches I and II given in Section 3.

Let us apply the Approach I. $J^T(y)$ can be represented by

$$J^T(y) = A_0 + \sum_{k=1}^4 y_k A_k, \quad (77)$$

where

$$A_0 = \begin{pmatrix} -2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & -2 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2 \\ & & & & -1 & 1 & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ & & & & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -2 & -2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & -4 & -4 & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 4 & -4 & -4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 & 0 \end{pmatrix}$$

It has been demonstrated in Section 3 that all simultaneously invariant subspaces for A_k ($k=0, 1, \dots, 4$) will give the full set of the fixed $J^T(y)$ -invariant ones and these simultaneously invariant subspaces are contained in $\text{Inv}(A)$, where

$$A = \sum_{k=0}^4 A_k \quad (78)$$

Using the method presented in Appendix B one can determine $\text{Inv}(A)$. We have

$$A = \begin{pmatrix} -4 & -2 & 2 & 2 & 1 & 0 & 1 & 0 \\ -2 & -4 & 2 & 2 & 2 & 0 & 0 & 0 \\ 4 & 4 & -6 & -4 & 0 & 2 & 0 & 0 \\ 4 & 4 & -4 & -6 & 0 & 0 & 0 & 2 \\ & & & & -1 & 1 & 0 & 0 \\ & 0 & & & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ & & & & 0 & 0 & 1 & -1 \end{pmatrix}.$$

The eigenvalues of A are $-14; -2, -2, -2; -1 - \sqrt{2}, -1 - \sqrt{2}; 0, 0$. The canonical form of $A - \lambda I_n$ (Appendix C) is the following:

$$\begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & \lambda + 2 & & \\ & & & & & & \lambda(\lambda + 2)(\lambda + 1 + \sqrt{2}) & \\ & & & & & & & \lambda(\lambda + 2)(\lambda + 1 + \sqrt{2})(\lambda + 14) \end{pmatrix}.$$

Notice that all the powers of the elementary divisors are unity, so the algebraic and geometric multiplicities of all the multiple eigenvalues are equal; the Jordan canonical form of A is a diagonal matrix and A has full eigenvectors. Each Jordan chain only has one vector. The root subspace for each eigenvalue is spanned by the corresponding eigenvectors. Arranging the eigenvectors according to the order of their eigenvalues given above, the eigenvector matrix X of A is the following:

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 & \frac{35 - 23\sqrt{2}}{167} & \frac{218 + 81\sqrt{2}}{167} & \frac{3}{7} & \frac{4}{7} \\ 1 & 0 & 0 & 1 & \frac{-132 - 190\sqrt{2}}{167} & \frac{-116 - 86\sqrt{2}}{167} & \frac{13}{14} & \frac{1}{14} \\ -2 & 0 & 1 & 0 & \frac{264 + 46\sqrt{2}}{167} & \frac{232 + 172\sqrt{2}}{167} & \frac{8}{7} & \frac{-1}{7} \\ -2 & 1 & 0 & 1 & \frac{-404 - 288\sqrt{2}}{167} & \frac{-102 - 162\sqrt{2}}{167} & \frac{1}{7} & \frac{6}{7} \\ & & & & 1 & 0 & 1 & 0 \\ 0 & & & & -\sqrt{2} & 0 & 1 & 0 \\ & & & & 0 & -\sqrt{2} & 0 & 1 \\ & & & & 0 & 1 & 0 & 1 \end{pmatrix}.$$

In this case, any linear combination of the eigenvectors for each multiple eigenvalue is still an eigenvector of A and spans a 1-dimensional invariant subspace of A . Also any two linearly independent such combinations for eigenvalue -2 span a 2-dimensional A -invariant subspace.

According to the relation between the invariant subspaces and the root subspaces, any A -invariant subspace with a given dimension can only be either an

invariant subspace in a root subspace or a sum of several lower dimensional invariant subspaces from different root subspaces. All invariant subspaces in a

root subspace can be easily determined, and their combinations will give all A -invariant subspaces. For the sake of brevity, we use x_i to represent column i of X . The 1-dimensional A -invariant subspaces are as follows:

$$\text{Span}\{0\}, \text{Span}\{x_1\}, \text{Span}\{x_1x_2 + x_2x_3 + x_3x_4\},$$

$$\text{Span}\{x_1x_5 + x_2x_6\}, \text{Span}\{x_1x_7 + x_2x_8\},$$

where $\alpha_i \in \mathcal{A}$ (the field of real numbers). Similarly we have all 2-dimensional A -invariant subspaces:

$$\text{Span}\{x_1, x_1x_2 + x_2x_3 + x_3x_4\},$$

$$\text{Span}\{x_1, x_1x_5 + x_2x_6\},$$

$$\text{Span}\{x_1, x_1x_7 + x_2x_8\},$$

$$\text{Span}\{x_1x_2 + x_2x_3 + x_3x_4, \beta_1x_2 + \beta_2x_3 + \beta_3x_4\},$$

$$\text{Span}\{x_1x_2 + x_2x_3 + x_3x_4, \beta_1x_5 + \beta_2x_6\},$$

$$\text{Span}\{x_1x_2 + x_2x_3 + x_3x_4, \beta_1x_7 + \beta_2x_8\},$$

$$\text{Span}\{x_5, x_6\}, \text{Span}\{x_1x_5 + x_2x_6, \beta_1x_7 + \beta_2x_8\},$$

$$\text{Span}\{x_7, x_8\},$$

where $\alpha_i, \beta_i \in \mathcal{A}$ and if a subspace contains the same number of α_i 's and β_i 's, the vectors α and β are linearly independent. In the same way we can determine all other A -invariant subspaces of dimensions higher than 2. To save space we will not list all elements of $\text{Inv}(A)$.

After examining which subspaces in $\text{Inv}(A)$ are simultaneously invariant under all A_k 's, we obtained 23 distinct types of fixed $J^T(y)$ -invariant subspaces.

According to the results in Section 3 these subspaces compose the full set of the fixed $J^T(y)$ -invariant subspaces. Choosing some bases for the invariant subspaces \mathcal{H} the corresponding matrices M are as follows. For other bases M can have different forms.

The matrices for 1-dimensional \mathcal{H} :

$$M_1 = (x_1 + x_2 \quad x_3 \quad x_2 \quad x_1 + x_3 \quad 0 \quad 0 \quad 0 \quad 0),$$

$$M_2 = (2 \quad -2\sqrt{2} \quad 4 \quad -2-2\sqrt{2} \quad 2-\sqrt{2} \quad 2-\sqrt{2} \quad 2-2\sqrt{2} \quad -\sqrt{2} \quad 1),$$

$$M_3 = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1).$$

Here the subspace spanned by the row vector of M_2 belongs to $\text{Span}\{x_1x_5 + x_2x_6\}$, and the subspace spanned by the row vector of M_3 belongs to $\text{Span}\{x_1x_7 + x_2x_8\}$. Note that only when x_1 and x_2 take on special values (for M_2 , $x_1 = 2 - \sqrt{2}$, $x_2 = 1$; for M_3 , $x_1 = x_2 = 1$) will the subspaces belonging to $\text{Span}\{x_1x_5 + x_2x_6\}$ and $\text{Span}\{x_1x_7 + x_2x_8\}$ be $J^T(y)$ -invariant. For M_4 there are 3 linearly independent row vectors according to the different values the α_i 's can take.

The matrices for 2-dimensional \mathcal{H} :

$$M_4 = \begin{pmatrix} x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \\ \beta_1 + \beta_2 & \beta_3 & \beta_2 & \beta_1 + \beta_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_5 = \begin{pmatrix} 2 & -2\sqrt{2} & 4 & -2-2\sqrt{2} & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \\ x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & -2\sqrt{2} & 4 & -2-2\sqrt{2} & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \end{pmatrix}.$$

The matrices for 3-dimensional \mathcal{H} :

$$M_8 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_9 = \begin{pmatrix} 2 & -2\sqrt{2} & 4 & -2-2\sqrt{2} & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \\ x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \\ \beta_1 + \beta_2 & \beta_3 & \beta_2 & \beta_1 + \beta_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_{10} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \\ \beta_1 + \beta_2 & \beta_3 & \beta_2 & \beta_1 + \beta_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_{11} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & -2\sqrt{2} & 4 & -2-2\sqrt{2} & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \\ x_1 + x_2 & x_3 & x_2 & x_1 + x_3 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrices for 4-dimensional \mathcal{M} :

$$M_{12} = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & 0 \\ & & & 1 \end{pmatrix},$$

$$M_{13} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix},$$

$$M_{14} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \end{pmatrix},$$

$$M_{15} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & -2\sqrt{2} & 4 & -2-2\sqrt{2} & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \\ \alpha_1+\alpha_2 & \alpha_3 & \alpha_2 & \alpha_1+\alpha_3 & 0 & 0 & 0 & 0 \\ \beta_1+\beta_2 & \beta_3 & \beta_2 & \beta_1+\beta_3 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

$$M_{20} = \begin{pmatrix} 1 & & & & & \\ & 1 & & 0 & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & 1 & 0 & 0 \\ & & & & 0 & 0 & 1 & 1 \end{pmatrix}.$$

The matrices for 5-dimensional \mathcal{M} :

$$M_{16} = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & & & 1 & \\ & & & & \alpha & -\sqrt{2}\alpha & -\sqrt{2}\beta & \beta \end{pmatrix},$$

$$M_{17} = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & & & 1 & \\ & & & & \gamma & \gamma & \delta & \delta \end{pmatrix},$$

$$M_{18} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 2-\sqrt{2} & 2-2\sqrt{2} & -\sqrt{2} & 1 \end{pmatrix}.$$

$$M_{21} = \begin{pmatrix} 1 & & & & \\ & 1 & & 0 & \\ & & 1 & & \\ & & & 1 & \\ & & & & \alpha & -\sqrt{2}\alpha & -\sqrt{2}\beta & \beta \\ & & & & \gamma & \gamma & \delta & \delta \end{pmatrix}.$$

The matrices for 7-dimensional \mathcal{M} :

$$M_{22} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & 0 & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & 1 \\ & & & & \gamma & \gamma & \delta & \delta \end{pmatrix},$$

The matrices for 6-dimensional \mathcal{M} :

$$M_{19} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & 1 \end{pmatrix},$$

$$M_{23} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & 0 & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & 1 & 0 & 0 \\ & & & & 0 & 0 & 1 & 1 \\ & & & & \alpha & -\sqrt{2}\alpha & -\sqrt{2}\beta & \beta \end{pmatrix}.$$

$\alpha_i, \beta_i, \alpha, \beta, \gamma, \delta \in \mathcal{A}$. If a matrix contains the same number of α_i 's and β_i 's, the vectors α and β are linearly independent.

By definition, a set S of subspaces of R^n compose a lattice if and only if $\{0\}$ and R^n belong to S and in addition S contains the intersection and sum of any

$$\begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & \lambda+2 \end{pmatrix} \begin{matrix} \\ \\ \\ \\ \\ \lambda(\lambda+2)(\lambda+1+\sqrt{2}) \\ \lambda(\lambda+2)(\lambda+1+\sqrt{2})(\lambda-\lambda(\mathbf{y})) \end{matrix}$$

two subspaces belonging to S (Gohberg *et al.*, 1986). We can demonstrate that all the fixed $J^T(\mathbf{y})$ -invariant subspaces with $\{0\}$ and R^n compose a lattice. Let \mathcal{H}' , \mathcal{H}'' be any two fixed $J^T(\mathbf{y})$ -invariant subspaces. If $\mathbf{x} \in \mathcal{H}' \cap \mathcal{H}''$, we have $J^T(\mathbf{y})\mathbf{x} \in \mathcal{H}'$ and $J^T(\mathbf{y})\mathbf{x} \in \mathcal{H}''$, so $J^T(\mathbf{y})\mathbf{x} \in \mathcal{H}' \cap \mathcal{H}''$ and $\mathcal{H}' \cap \mathcal{H}''$ is $J^T(\mathbf{y})$ -invariant. Now let $\mathbf{x} \in \mathcal{H}' + \mathcal{H}''$, so that $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$, where $\mathbf{x}_1 \in \mathcal{H}'$, $\mathbf{x}_2 \in \mathcal{H}''$. Then $J^T(\mathbf{y})\mathbf{x} = J^T(\mathbf{y})\mathbf{x}_1 + J^T(\mathbf{y})\mathbf{x}_2 \in \mathcal{H}' + \mathcal{H}''$. Therefore, $\mathcal{H}' + \mathcal{H}''$ is $J^T(\mathbf{y})$ -invariant as well. In accordance with the definition of a lattice, all the fixed $J^T(\mathbf{y})$ -invariant subspaces with $\{0\}$ and R^n compose a lattice. This conclusion is easy to check for all fixed $J^T(\mathbf{y})$ -invariant subspaces corresponding to the M_i 's given above.

This property has some utility here. We will find that some fixed $J^T(\mathbf{y})$ -invariant subspaces are irreducible. Here an irreducible invariant subspace \mathcal{H} of $J^T(\mathbf{y})$ means that it cannot be represented as a direct sum of nonzero $J^T(\mathbf{y})$ -invariant subspaces \mathcal{H}' and \mathcal{H}'' ; otherwise \mathcal{H} is called reducible. Let \mathcal{H}_i represent the subspace spanned by the row vectors of M_i . For the present problem the following fixed $J^T(\mathbf{y})$ -invariant subspaces are irreducible:

dimension 1: $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$;

dimension 4: \mathcal{H}_{12} ;

dimension 5: $\mathcal{H}_{16}, \mathcal{H}_{17}$;

dimension 6: $\mathcal{H}_{19}, \mathcal{H}_{20}, \mathcal{H}_{21}$;

dimension 7: $\mathcal{H}_{22}, \mathcal{H}_{23}$.

Other reducible ones can be obtained from the irreducible fixed $J^T(\mathbf{y})$ -invariant subspaces. We can also find that some fixed $J^T(\mathbf{y})$ -invariant subspaces are contained in other ones. There are also some chains in the fixed $J^T(\mathbf{y})$ -invariant subspace. One of them is the following:

$$\{0\} \in \mathcal{H}_1 \in \mathcal{H}_6 \in \mathcal{H}_{10} \in \mathcal{H}_{13} \in \mathcal{H}_{17} \in \mathcal{H}_{21} \in \mathcal{H}_{23} \in \mathcal{R}^n.$$

The above property of all $J^T(\mathbf{y})$ -invariant subspaces is not closely related to the present analysis. However, it does have significance in the study of other applications.

For the purposes of illustration we apply the Approach II in Section 3 to determine the corresponding matrices of the fixed $J^T(\mathbf{y})$ -invariant subspaces. The eigenvalues of $J^T(\mathbf{y})$ are $-2-2y_1-2y_2-4y_3-4y_4$; $-2, -2, -2$; $-1-\sqrt{2}$; $-1-\sqrt{2}$; $0, 0$. The canonical form of $J^T(\mathbf{y}) - \lambda I_n$ (Appendix C) is as follows:

where $\lambda(\mathbf{y}) = -2-2y_1-2y_2-4y_3-4y_4$. Notice that all the powers of the elementary divisors are unity and the minimal polynomial is the product of the polynomials with degree 1 for all distinct eigenvalues.

Solving the equation

$$[J^T(\mathbf{y}) - \lambda_i I_n] \mathbf{r}_i \mathbf{x} = 0, \quad (79)$$

we find that for any $r_i > 1$ the solutions consisting of constant vectors are the same as those of $r_i = 1$. Therefore, only $r_i = 1$ is considered. The results obtained are as follows:

$\lambda_i = -2$: M_8 containing M_1 and M_4 ;

$\lambda_i = -1-\sqrt{2}$: M_2 ;

$\lambda_i = 0$: M_3 .

Solving eq. (79) for $\lambda_i = -2$ gives three linearly independent constant solutions of \mathbf{x} , which are the basis of \mathcal{H}_8 . Note that the subspaces spanned by the row vectors of M_1 and M_4 , respectively, are subspaces of \mathcal{H}_8 . Therefore, when M_8 is given, M_1 and M_4 are also obtained. The same situation appears in the following results. These results also show that the corresponding invariant subspaces are associated with the constant eigenvalues.

Similarly solving the equation

$$[J^T(\mathbf{y}) - \lambda_i I_n][J^T(\mathbf{y}) - \lambda_j I_n] \mathbf{x} = 0, \quad (80)$$

we obtain

$\lambda_i = -2-2y_1-2y_2-4y_3-4y_4$, $\lambda_j = -2$: M_{12} ;

$\lambda_i = -2$, $\lambda_j = -1-\sqrt{2}$: M_{14} containing M_8 and M_9 ;

$\lambda_i = -2$, $\lambda_j = 0$: M_{13} containing M_8 and M_{10} ;

$\lambda_i = -1-\sqrt{2}$, $\lambda_j = 0$: M_{11} .

Solving the equation

$$[J^T(\mathbf{y}) - \lambda_i I_n][J^T(\mathbf{y}) - \lambda_j I_n][J^T(\mathbf{y}) - \lambda_k I_n] \mathbf{x} = 0 \quad (81)$$

gives

$\lambda_i = -2-2y_1-2y_2-4y_3-4y_4$, $\lambda_j = -2$, $\lambda_k = -1-\sqrt{2}$: M_{16} containing M_{14} ;

$\lambda_i = -2 - 2y_1 - 2y_2 - 4y_3 - 4y_4$, $\lambda_j = -2$, $\lambda_k = 0$: M_{20} containing M_{17} ;

$\lambda_i = -2$, $\lambda_j = -1 - \sqrt{2}$, $\lambda_k = 0$: M_{18} containing M_{11} and M_{15} .

Until now we have determined all M_j except M_{21} , M_{22} and M_{23} . We cannot determine them by solving the following equation containing all distinct eigenvalues

$$[J^T(y) - \lambda_1 I_n][J^T(y) - \lambda_2 I_n][J^T(y) - \lambda_3 I_n][J^T(y) - \lambda_4 I_n]x = 0, \quad (82)$$

because the left side of eq. (82) is associated with the minimal polynomial of $J^T(y)$ and becomes the null matrix. However, notice that the vectors orthogonal to all row vectors of M_{23} and M_{22} are

$$v_1 = (0 \ 0 \ 0 \ 0 \ \beta \ -\beta \ \alpha \ -\alpha)^T, \\ v_2 = (0 \ 0 \ 0 \ 0 \ \sqrt{2}\delta \ \delta \ -\gamma \ -\sqrt{2}\gamma)^T,$$

respectively. They can be respectively obtained by solving the equation

$$[J(y) - \lambda_i I_n]x = 0 \quad (83)$$

for eigenvalues $-1 - \sqrt{2}$ and 0. Similarly, the column vectors of the following matrix V are orthogonal to all row vectors of M_{21} :

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 & \beta & -\beta & \alpha & -\alpha \\ 0 & 0 & 0 & 0 & \sqrt{2}\delta & \delta & -\gamma & -\sqrt{2}\gamma \end{pmatrix}^T.$$

These two column vectors can be obtained by solving

$$[J(y) - \lambda_i I_n][J(y) - \lambda_j I_n]x = 0 \quad (84)$$

for $\lambda_i = -1 - \sqrt{2}$ and $\lambda_j = 0$. After they are determined, M_{23} , M_{22} and M_{21} can be obtained from their orthogonal complementary subspaces. Now all M_i obtained by the Approach I are also completely determined by the Approach II presented in Section 3.

From Section 2 we know that for nonlinear systems only some of these fixed $J^T(y)$ -invariant subspaces can be used to construct the lumping matrices. The remaining task is to examine which of them satisfy the sufficient condition for exact lumping. Examining M_i to M_{23} we can see that except for M_j ($j = 12, 16, 17, 19, 20-23$) all other matrices M_i are related to constant eigenvalues, and therefore they can be used as lumping matrices.

Let us consider M_j further. They have a common form as

$$M_j = \begin{pmatrix} I_4 & 0 \\ 0 & B \end{pmatrix}.$$

The generalized inverse of M_j should be of the form

$$\bar{M}_j = \begin{pmatrix} I_4 & 0 \\ 0 & \bar{B} \end{pmatrix}.$$

where $B\bar{B} = I_4$. Then we have

$$\bar{M}_j M_j = \begin{pmatrix} I_4 & 0 \\ 0 & \bar{B}B \end{pmatrix}$$

This shows that y_1, y_2, y_3 and y_4 do not change for $\bar{M}_j M_j y$. Then the eigenvalue $-2 - 2y_1 - 2y_2 - 4y_3 - 4y_4$ will be the same for $J^T(y)$ and $J^T(\bar{M}_j M_j y)$, and M_j can be also used for exact lumping. Thus all 23 distinct types of matrices M are lumping matrices.

Substituting any M into eq. (11) and arbitrarily choosing two different generalized inverses \bar{M} , we obtain the same differential equations for the lumped model associated with M . When the corresponding eigenvalues of M are constant, say for $M_1, M_2, M_3, M_{13}, M_{15}$, the lumped systems are linear. Their differential equations are, respectively, as follows:

$$d\hat{y}/dt = -2\hat{y}, \quad (85)$$

$$d\hat{y}/dt = -(1 + \sqrt{2})\hat{y}, \quad (86)$$

$$d\hat{y}/dt = 0, \quad (87)$$

$$\frac{d}{dt} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \end{pmatrix} = \begin{pmatrix} -2 & & & \\ 0 & -2 & & \\ 0 & 0 & -2 & \\ 0 & 2 & 2 & 0 \end{pmatrix} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \end{pmatrix}, \quad (88)$$

$$\frac{d}{dt} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \end{pmatrix} = \begin{pmatrix} 0 & & & \\ -(1 + \sqrt{2}) & & & \\ & -2 & & \\ & & -2 & \end{pmatrix} \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \hat{y}_4 \end{pmatrix}. \quad (89)$$

When the corresponding eigenvalue spectrum of M contains $-2 - 2y_1 - 2y_2 - 4y_3 - 4y_4$, say for M_{19} , the lumped model is no longer linear:

$$\begin{aligned} d\hat{y}_1/dt &= -2\hat{y}_1 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\ d\hat{y}_2/dt &= -2\hat{y}_2 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\ d\hat{y}_3/dt &= -2\hat{y}_3 + 4\hat{y}_3\hat{y}_4 + 2\hat{y}_1\hat{y}_2 \\ d\hat{y}_4/dt &= -2\hat{y}_4 - 4\hat{y}_3\hat{y}_4 + 2\hat{y}_1\hat{y}_2 \\ d\hat{y}_5/dt &= \hat{y}_1 - 2\hat{y}_2 - 2\sqrt{2}\hat{y}_3 + (1 + \sqrt{2})\hat{y}_5 \\ d\hat{y}_6/dt &= -\sqrt{2}\hat{y}_1 + 2\hat{y}_4 - (1 + \sqrt{2})\hat{y}_6 \end{aligned} \quad (90)$$

where

$$\hat{y}_i = y_i, \quad (i = 1, 2, 3, 4)$$

$$\hat{y}_5 = y_5 - \sqrt{2}y_6,$$

$$\hat{y}_6 = -\sqrt{2}y_5 + y_6.$$

We have obtained 23 distinct kinds of lumping matrices. Actually there are an infinite number of lumping matrices, if we give different values for parameters α, β, γ and δ . We can also construct other lumping matrices by elementary row operations on the rows of M . For example, letting $\gamma = \delta = 1$ for M_1 , we obtain

$$M_1 = \begin{pmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & 1 \\ & & & & & 1 \end{pmatrix}.$$

Similarly letting $x_1=0$, $x_2=x_3=1$ and using elementary row operations on the two rows of M_6 gives

$$M'_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Letting $\alpha=0$, $\beta=1$ or $\alpha=1$, $\beta=0$ and using different elementary row operations on the last 3 rows of M_{23} , we have

$$M'_{23} = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$M''_{23} = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

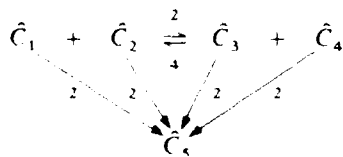
These special cases have a particular significance argued below. Usually the lumped model of a uni- and/or bimolecular reaction system does not follow a uni- and/or bimolecular reaction scheme. However, there is a special group of lumping matrices called "proper lumping matrices" (Wei and Kuo, 1969), each column of which is a unit vector e_i . It has been proved (Wei and Kuo, 1969; Li, 1984) that for proper lumping the lumped model follows a uni- and/or bimolecular reaction scheme. In Example 2 there are some proper lumping matrices, such as M'_6 , M'_{17} , M_{20} , M'_{23} and M''_{23} . The corresponding lumped models are as follows:

lumping matrix M'_6 :



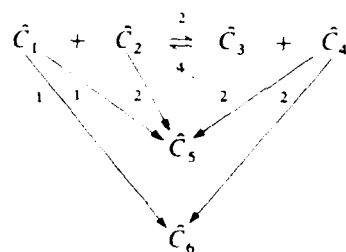
$$\hat{C}_1 = \sum_{i=1}^4 C_i, \quad \hat{C}_2 = \sum_{i=5}^8 C_i.$$

lumping matrix M'_{17} :



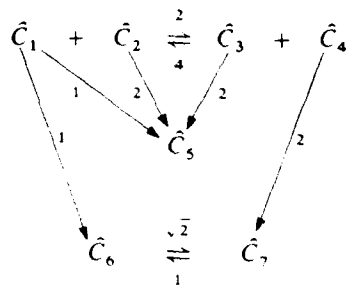
$$\hat{C}_i = C_i \quad (i=1, 2, 3, 4), \quad \hat{C}_5 = \sum_{i=5}^8 C_i.$$

lumping matrix M_{20} :



$$\hat{C}_i = C_i \quad (i=1, 2, 3, 4), \quad \hat{C}_5 = C_5 + C_6, \quad \hat{C}_6 = C_7 + C_8.$$

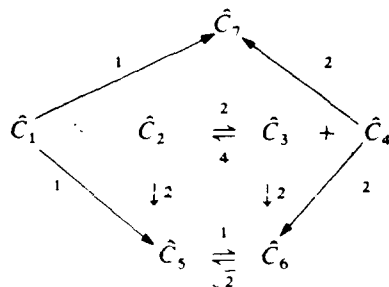
lumping matrix M'_{23} :



$$\hat{C}_i = C_i \quad (i=1, 2, 3, 4), \quad \hat{C}_i = C_{i+1} \quad (i=6, 7),$$

$$\hat{C}_5 = C_5 + C_6.$$

lumping matrix M''_{23} :



$$\hat{C}_i = C_i \quad (i=1, 2, 3, 4, 5, 6), \quad \hat{C}_7 = C_7 + C_8.$$

To summarize, by these two examples we have illustrated how to apply the methods to determine all the lumping matrices. First we need to determine all the fixed $J^T(y)$ -invariant subspaces. There are two approaches to achieve this task. One is associated with the decomposition of $J^T(y)$ into a linear combination of some basis constant matrices and the subsequent determination of the simultaneously invariant subspaces for all these constant matrices; the other one is dependent on the determination of the fixed null subspaces of the different products of the λ -matrices $J^T(y) - \lambda_i I_n$ for all distinct eigenvalues. After the determination of all the fixed $J^T(y)$ -invariant subspaces, we need to examine which of them satisfy the sufficient condition for exact lumping and then we use these subspaces to construct lumping matrices. The results show that for uni- and/or bimolecular reaction systems one can determine all possible lumping matrices. These examples are very simple, however, they illustrate the methods which can be applied to other more complicated systems.

5. CONCLUSION AND DISCUSSION

In this paper a general analysis of exact lumping has been given, which can be used for any system described by a set of first order ordinary differential equations with any degree of nonlinearity. Uni- and/or bimolecular reaction systems are only special cases of this general analysis.

A systematic method to determine all the fixed invariant subspaces for the transpose of the Jacobian matrix of the kinetic equations and all the lumping matrices was developed. Using the generalized inverse of the lumping matrix, the differential equations of the lumped system can be readily obtained, and the non-unique nature of the generalized inverses does not effect the form of the lumped equations in the exact case.

In the present work lumping is considered to be generated by a linear transformation. In spite of a system being nonlinear, this paper shows that under appropriate conditions linear transformation can still lead to exact lumping. If a nonlinear system is exactly lumpable in this sense, it must possess a degree of partial linearity. Therefore, it is natural that the lumpability of a nonlinear system is related to some fixed invariant subspaces and the invariance of the corresponding eigenvalues for the transpose of the Jacobian matrix. The partial linearity of nonlinear systems is useful not only for simplification of a complicated system, but it also provides physical insight. For example, eq. (87) shows that the fixed invariant subspace spanned by the row vector M_3 is connected with the property of mass conservation. Using the same approach for classical mechanics systems we could yield other conservation properties.

Although some useful results about exact lumping have been obtained, there is still further work to do. Systematic application of this analysis to complex reaction systems needs to be considered. However, in the treatment of actual reaction systems, the first problem encountered will likely be their non-exact lumpability. Sometimes, even if a system is exactly lumpable, the results may not meet practically desired goals. For example, in the CO/H₂O/O₂ combustion system we would like the easily measurable concentrations of CO, CO₂, O₂, H₂O to be unlumped. With this constraint, the system likely can not be exactly lumped, and we have to lump the other radical species of the system approximately. Developing a general approach for approximate lumping is very important for realistic problems. The exact lumping analysis presented above should form a rigorous starting point for the development of approximate lumping.

Acknowledgement The authors acknowledge support from the Air Force Office of scientific research.

NOTATION

Scalars

a_i defined as $\sum_{k=1}^{p_i} c_{ki}$

$a_k(y)$	k th coefficient of a linear combination of constant matrices for $J^T(y)$
C_i	i th species of a reaction system
\bar{C}_i	i th species of a lumped system
c	constant
$d_k(\lambda)$	k th invariant polynomial
e_{kj}	partial multiplicity
$f_i(y)$	i th element of $f(y)$
$\text{Inv}(A)$	set of all A -invariant subspaces
i	positive integer
$j_{ij}(y)$	(i, j) -entry of matrix $J(y)$
$\text{Ker } A'$	null subspace of A'
l	positive integer
\mathcal{H}	invariant subspace of $J^T(y)$ or A
\mathcal{V}	null subspace of M
n	dimension of vector y
\hat{n}	dimension of vector \hat{y}
p_i	minimal value of positive integer r_i for the largest $\text{Ker}(A - \lambda_i I)^{r_i}$
q_j	minimal value of positive integer r_j for the largest $\text{Ker}[(\sigma_j^2 + \tau_j^2)I - 2\sigma_j J^T(y) + (J^T(y))^2]^{r_j}$
q_{ij}	(i, j) -entry of $Q(y)$
\mathcal{R}	field of real number
$\mathcal{R}_{\lambda_i}(A)$	root subspace for real eigenvalue λ_i of A
$\mathcal{R}_{\sigma_i \pm i\tau_i}(A)$	root subspace for nonreal eigenvalues $\sigma_i \pm i\tau_i$ of A
\mathcal{R}^n	n -dimensional real space
r	nonnegative integer
r_i	nonnegative integer
S	set of invariant subspaces
s	positive integer
t	time or positive integer
y_k	k th element of vector y

Vectors and matrices

Capital letters represent matrices; bold-face lower case letters represent vectors.

A	constant matrix
A_0	constant matrix
A_k	constant matrix
$A(y)$	2×2 function matrix
B	matrix
\bar{B}	generalized inverse of B
c	\hat{n} -dimensional arbitrary constant vector
E_{ij}	elementary matrix with 1 as its (i, j) -entry, and 0 for the rest of the elements
e_i	unit vector with 1 as its i th element, and 0 for the rest of the elements
$f(y)$	n -dimensional function vector
$\hat{f}(\hat{y})$	\hat{n} -dimensional function vector
I	identity matrix
$J(y)$	Jacobian matrix of $f(y)$
$J^T(y)$	transpose of Jacobian matrix
$J_{or}(\lambda)$	Jordan matrix
$J_{or}[\lambda(y)]$	Jordan matrix
$J_{or}^p(\lambda_i)$	Jordan block for real eigenvalue λ_i
$J_{or}^q(\sigma_j \pm i\tau_j)$	Jordan block for nonreal eigenvalues $\sigma_j \pm i\tau_j$
K	rate constant matrix
K_j	2×2 submatrix

M	lumping matrix
M_{sq}	nonsingular $\hat{n} \times \hat{n}$ submatrix of M
$m_{(i)}$	transpose of row i of lumping matrix M
\bar{M}	generalized inverse of M satisfying $M\bar{M} = I_n$
Q	$\hat{n} \times \hat{n}$ constant matrix
$Q(y)$	$\hat{n} \times \hat{n}$ function matrix
$S(y)$	$\hat{n} \times \hat{n}$ invertible matrix
V	8×2 constant matrix
v_i	8-dimensional constant vector
x	n -dimensional vector
X	eigenvector matrix
y	n -dimensional variable vector
\hat{y}	\hat{n} -dimensional variable vector

Greek letters

α	real number
β	real number
γ	real number
δ	real number
λ	eigenvalue of a matrix
λ_i	i th eigenvalue of a matrix
$\lambda(y)$	eigenvalue vector
Λ	diagonal eigenvalue matrix of K with λ_i as its i th diagonal element
σ	real number
τ	real number

Symbols

\wedge	any property related to the lumped system
0	null vector
0	null matrix

REFERENCES

- Bailey, J. E., 1972, Lumping analysis of reactions in continuous mixtures. *Chem. Engng J.* **3**, 52-61.
 Bailey, J. E., 1975, Diffusion of grouped multicomponent mixtures in uniform and nonuniform media. *A.I.Ch.E. J.* **21**, 192-194.

- Hutchinson, P. and Luss, D., 1970, Lumping of mixtures with many parallel first order reactions. *Chem. Engng J.* **1**, 129-135.
 Isral, A. B. and Greville, T. N. E., 1974, *Generalized Inverse: Theory and Applications*. Wiley, New York.
 Jacob, S. M., Gross, B., Voltz, S. E. and Weekman, V. W., Jr., 1976, A lumping and reaction scheme for catalytic cracking. *A.I.Ch.E. J.* **22**, 701-713.
 Lang, S., 1986, *Introduction to Linear Algebra*, 2nd edition. Springer, New York.
 Li, G., 1984, A lumping analysis in mono- or/and bimolecular reaction systems. *Chem. Engng Sci.* **39**, 1261-1270.
 Luss, D. and Hutchinson, P., 1971, Lumping of mixture with many parallel N -th order reactions. *Chem. Engng J.* **2**, 172-177.
 Luss, D., 1975, Grouping of many species each consumed by two parallel first-order reactions. *A.I.Ch.E. J.* **21**, 865-872.
 Ozawa, Y., 1973, The structure of a lumpable monomolecular system for reversible chemical reactions. *Ind. Engng Chem. Fundam.* **12**, 191-196.
 Wei, J. and Kuo, J. C. W., 1969, A lumping analysis in monomolecular reaction systems. *Ind. Engng Chem. Fundam.* **8**, 114-133.
 Wei, J. and Prater, C. D., 1963, A new approach to first-order chemical reaction systems. *A.I.Ch.E. J.* **9**, 77-81.

APPENDICES

The material in these Appendices concerns certain matrix operations and properties, particularly relevant to this paper. Although this material may be found in the literature, we present it here for completeness and convenience of the reader.

Appendix A: Jordan form of an $n \times n$ real matrix

In chemical kinetics we usually treat real matrices, and therefore, in the appendices we only deal with them. All the results in the appendices can be directly extended to treat nonreal matrices (Gohberg *et al.*, 1986).

Let A be an $n \times n$ real matrix. All distinct eigenvalues of A are $\lambda_1, \lambda_2, \dots, \lambda_t, \sigma_1 \pm i\tau_1, \sigma_2 \pm i\tau_2, \dots, \sigma_s \pm i\tau_s$. Each one may have multiplicity higher than 1. Here, λ_i, σ_i and τ_i are real numbers and τ_i are positive. For a real matrix nonreal eigenvalues appear in complex conjugate pairs. There exists a real similarity transformation matrix S such that

$$S^{-1}AS = J_{or}(\lambda), \quad (A1)$$

with $J_{or}(\lambda)$ being the Jordan matrix of the form

$$J_{or}(\lambda) = \begin{pmatrix} J_{or}^1(\lambda_1) & & & & \\ & J_{or}^2(\lambda_1) & & & \\ & & \ddots & & \\ & & & J_{or}^{p_1}(\lambda_1) & \\ & & & & J_{or}^1(\lambda_2) \\ & & & & & \ddots \\ & & & & & & J_{or}^{p_t}(\lambda_t) \\ & & & & & & & J_{or}^1(\sigma_1, \tau_1) \\ & & & & & & & & \ddots \\ & & & & & & & & & J_{or}^{q_s}(\sigma_s, \tau_s) \end{pmatrix}, \quad (A2)$$

where $J_{or}^p(\lambda_i), J_{or}^q(\sigma_j, \tau_j)$ ($p = 1, 2, \dots, p_i; q = 1, 2, \dots, q_j$) are called Jordan blocks with the following forms, respectively. The meaning of p_i and q_j is given below.

$$J_{or}^p(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 \\ 0 & 0 & \dots & \dots & \lambda_i \end{pmatrix}, \quad (A3)$$

- Golikeri, S. V. and Luss, D., 1972, Analysis of activation energy of grouped parallel reactions. *A.I.Ch.E. J.* **18**, 277-282.
 Golikeri, S. V. and Luss, D., 1974, Aggregation of many coupled consecutive first order reactions. *Chem. Engng Sci.* **29**, 845-855.
 Graham, A., 1981, *Knonecker Products and Matrix Calculus: with Applications*. Ellis Horwood, New York.
 Gohberg, I., Lancaster, P. and Rodman, L., 1986, *Invariant Subspaces of Matrices with Applications*. Wiley, New York.

$$J_{or}^q(\sigma_j, \tau_j) = \begin{pmatrix} K_j & I_2 & 0 & \dots & 0 \\ 0 & K_j & I_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & I_2 \\ 0 & 0 & \dots & \dots & K_j \end{pmatrix}, \quad (A4)$$

where I_2 represents the 2×2 identity matrix and

$$K_j = \begin{pmatrix} \sigma_j & \tau_j \\ -\tau_j & \sigma_j \end{pmatrix}. \quad (A5)$$

In the expression (A2) the blocks $J_{or}^p(\lambda_i)$ and $J_{or}^q(\sigma_j, \tau_j)$ are uniquely determined by A up to a permutation of their ordering.

Let $J_{or}^1(\lambda_i), \dots, J_{or}^{p_i}(\lambda_i)$ be all the Jordan blocks in expression (A2) for eigenvalue λ_i of A . The positive integer p_i is called the geometric multiplicity of λ_i . The dimension of each Jordan block is called the partial multiplicity and the sum of all partial multiplicities for λ_i is its algebraic multiplicity. The partial multiplicities of the Jordan blocks corresponding to the nonreal eigenvalue $\sigma_j + i\tau_j$ (or $\sigma_j - i\tau_j$) of A are, by definition, the half-sizes of the blocks $J_{or}^q(\sigma_j, \tau_j)$. The number of the blocks corresponding to (σ_j, τ_j) is the geometric multiplicity of $\sigma_j + i\tau_j$ (or $\sigma_j - i\tau_j$), and the sum of all partial multiplicities for $(\sigma_j + i\tau_j)$ (or $\sigma_j - i\tau_j$) is its algebraic multiplicity. Obviously, the algebraic multiplicity of an eigenvalue is not less than its geometric multiplicity. When all partial multiplicities are equal to unity, the algebraic and geometric multiplicities for each eigenvalue are equal, and in addition, when all eigenvalues are real, the Jordan matrix becomes diagonal with the eigenvalues as its diagonal elements. In this case S is the eigenvector matrix of A .

Appendix B: $\text{Inv}(A)$

The set of all invariant subspaces of a matrix B and the set of its similar matrix SBS^{-1} are related as

$$S[\text{Inv}(B)] = \text{Inv}(SBS^{-1}) \quad (B1)$$

with S being a similarity transformation matrix. Therefore, it is desirable to use similarity transformations to reduce a matrix to the simplest form for the determination of the set of all invariant subspaces. The "simplest form" here is the Jordan matrix. Let $B = J_{or}(\lambda)$, and eq. (B1) becomes

$$\begin{aligned} S[\text{Inv}(J_{or}(\lambda))] &= \text{Inv}[SJ_{or}(\lambda)S^{-1}] \\ &= \text{Inv}(A). \end{aligned} \quad (B2)$$

To determine $\text{Inv}(A)$ we need to determine $\text{Inv}[J_{or}(\lambda)]$.

First let us consider a set of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l$ such that

$$\begin{aligned} A\mathbf{x}_1 &= \lambda\mathbf{x}_1, \\ A\mathbf{x}_i &= \lambda\mathbf{x}_i + \mathbf{x}_{i-1}, \quad (i=2, 3, \dots, l) \end{aligned} \quad (B3)$$

We call \mathbf{x}_1 the eigenvector and \mathbf{x}_i ($i \geq 2$) generalized eigenvectors corresponding to eigenvalue λ and $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l$ are called a Jordan chain.

Without loss of generality we consider the first Jordan block in expression (A2).

$$J_{or}^1(\lambda_1) = \begin{pmatrix} \lambda_1 & 1 & 0 & \dots & 0 \\ 0 & \lambda_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & 1 \\ 0 & 0 & \dots & \dots & \lambda_1 \end{pmatrix}. \quad (B4)$$

Let the dimension of $J_{or}^1(\lambda_1)$ be e_{11} . Since

$$J_{or}^1(\lambda_1)\mathbf{e}_1 = \lambda_1\mathbf{e}_1,$$

$$J_{or}^1(\lambda_1)\mathbf{e}_i = \lambda_1\mathbf{e}_i + \mathbf{e}_{i-1}, \quad (i=2, 3, \dots, e_{11}) \quad (B5)$$

all subspaces spanned by a Jordan chain $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{e_{11}}$ are

$J_{or}^1(\lambda_1)$ -invariant. It can be also proved that any $J_{or}^1(\lambda_1)$ -invariant subspace is of the form $\text{Span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_i\}$. Therefore, $J_{or}^1(\lambda_1)$ only has e_{11} nonzero invariant subspaces. Similarly $J_{or}^2(\lambda_1), \dots, J_{or}^{p_1}(\lambda_1)$ have e_{21}, \dots, e_{p_11} invariant subspaces, respectively. Here e_{21}, \dots, e_{p_11} are the corresponding dimensions of the Jordan blocks for λ_1 . If we expand \mathbf{e}_i to n -dimensional vectors by adding zeroes at the end of each \mathbf{e}_i , all these subspaces are also $J_{or}(\lambda)$ -invariant. In addition, the sum of any set of the invariant subspaces corresponding to different Jordan blocks for λ_1 is also $J_{or}(\lambda)$ -invariant. Considering that there are p_1 eigenvectors corresponding to λ_1 , it follows that their linear combinations will give other eigenvectors for λ_1 . They compose an infinite number of 1-dimensional invariant subspaces, when $p_1 > 1$. All these considerations give the full group of $J_{or}(\lambda)$ -invariant subspaces corresponding to λ_1 . Let

$$a_1 = \sum_{k=1}^{p_1} e_{k1}. \quad (B6)$$

Then the biggest $J_{or}(\lambda)$ -invariant subspace corresponding to λ_1 is $\text{Span}\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{a_1}\}$, which is called the root subspace of $J_{or}(\lambda)$ corresponding to λ_1 and is denoted by $\mathcal{R}_{\lambda_1}[J_{or}(\lambda)]$. All $J_{or}(\lambda)$ -invariant subspaces considered above belong to it. Similarly we can construct all other $J_{or}(\lambda)$ -invariant subspaces belonging to $\mathcal{R}_{\lambda_i}[J_{or}(\lambda)]$, and all the sums of invariant subspaces belonging to different root subspaces give $\text{Inv}[J_{or}(\lambda)]$.

There is a one-to-one correspondence between $\text{Inv}[J_{or}(\lambda)]$ and $\text{Inv}(A)$. $\text{Inv}(A)$ can be constructed in the same way except for using eigenvectors and generalized eigenvectors of A instead of \mathbf{e}_i . If we know the eigenvalues and the corresponding Jordan form of A , its eigenvectors and generalized eigenvectors can be readily determined by solving eq. (B3). The Jordan form of A is easy to work out when we obtain the canonical form of the λ -matrix $A - \lambda I_n$. This will be discussed in Appendix C.

When A has nonreal eigenvalues and we are only interested in real invariant subspaces for a pair of conjugate nonreal eigenvalues, then the Jordan block is given by expression (A4) in Appendix A. The only difference is that any Jordan chain now has an even number of vectors: $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{2j}$ and each Jordan block contains a unique A -invariant subspace with dimension 2.

Appendix C: canonical form of λ -matrix $A - \lambda I_n$

$A - \lambda I_n$ is called the λ -matrix of A . Using elementary row and column operations $A - \lambda I_n$ can be transformed to its canonical form:

$$\begin{pmatrix} d_1(\lambda) & & & & \\ & d_2(\lambda) & & & \\ & & \ddots & & \\ & & & d_r(\lambda) & \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & 0 \end{pmatrix}, \quad (C1)$$

where r is the rank of $A - \lambda I_n$; $d_k(\lambda)$ are called invariant polynomials, which are polynomials of λ with the leading coefficient 1 and $d_k(\lambda)|d_{k+1}(\lambda)$ for $k=1, 2, \dots, r-1$. Here $d_k(\lambda)|d_{k+1}(\lambda)$ means that $d_{k+1}(\lambda)$ is divisible by $d_k(\lambda)$. Especially note that $d_r(\lambda)$ is called the minimal polynomial of A and satisfies $d_r(A)=0$.

Let $\lambda_1, \lambda_2, \dots, \lambda_t$ be all the distinct eigenvalues of A . Then $d_k(\lambda)$ can be further decomposed as

$$d_k(\lambda) = (\lambda - \lambda_1)^{e_{k1}}(\lambda - \lambda_2)^{e_{k2}} \dots (\lambda - \lambda_t)^{e_{kt}}, \quad (k=1, 2, \dots, r) \quad (C2)$$

Since we have the property $d_k(\lambda)|d_{k+1}(\lambda)$, then it follows that

$$e_{1j} \leq e_{2j} \leq \dots \leq e_{rj}, \quad (j=1, 2, \dots, t) \quad (C3)$$

Considering that $\lambda_1, \lambda_2, \dots, \lambda_t$ are distinct and $d_k(\lambda)|d_{k+1}(\lambda)$ for all k , so all e_{rj} are not equal to zero.

Appendix I

9. Determination of Constrained Lumping Schemes for Nonisothermal First-order Reaction Systems, G. Li and H. Rabitz, Chem Eng. Sci., 46, 583 (1991).

DETERMINATION OF CONSTRAINED LUMPING SCHEMES FOR NONISOTHERMAL FIRST-ORDER REACTION SYSTEMS

GENYUAN LI and HERSCHEL RABITZ*

Department of Chemistry, Princeton University, Princeton, NJ 08540, U.S.A.

(First received 22 January 1990; accepted in revised form 11 April 1990)

Abstract—The direct approach to determining the constrained lumping schemes presented in a previous paper is applied to nonisothermal first-order reaction systems. The constant basis matrices of the transpose of the Jacobian matrix for the kinetic equations are replaced by a set of rate constant matrices at different temperatures, which properly cover the desired temperature region. The Mobil "10-lump cracking model" is used as an example to illustrate this approach.

1. INTRODUCTION

Our previous paper (Li and Rabitz, 1991) presented a direct approach to determining the constrained lumping schemes for an arbitrary reaction system. When the system is isothermal, the transpose of the Jacobian matrix of the kinetic equations can be readily decomposed as a linear combination of a set of constant matrices. They are viewed as a basis of the transpose of the Jacobian matrix. Using the concept of the simultaneous minimal invariant subspace to all these basis matrices over a given subspace, the direct approach will supply the best constrained lumping matrices with different dimensions. For a nonisothermal first-order reaction system the transpose of the Jacobian matrix is the transpose of the rate constant matrix, which is a function of temperature and also has a set of constant basis matrices. Therefore, the direct approach can, in principle, be employed to determine the constrained lumping matrices for this system if one can find the basis matrices. Unfortunately, the rate constants are generally exponential functions of temperature and then it is not easy to determine the constant basis matrices of the transpose of the rate constant matrix. However, the basis matrices can simply be replaced by a set of rate constant matrices corresponding to different fixed temperatures in the desired temperature region. When the number of chosen constant matrices in the set is large enough and the temperature region is properly covered by the chosen temperature points, the results will be the same or close to those obtained by using the basis matrices. In Section 2 the theoretical basis of the direct approach for application to nonisothermal first-order reaction systems is presented. The Mobil "10-lump cracking model" is used as an example to illustrate this method in Section 3. Finally, Section 4 presents a conclusion and discussion.

2. THE DIRECT APPROACH FOR NONISOTHERMAL FIRST-ORDER REACTION SYSTEMS

Our previous papers (Li and Rabitz, 1989, 1990) presented a general analysis of exact and approximate lumping for a reaction system in a desired region Ω of the composition Y_n -space. The original reaction system with n -components can be described by

$$dy/dt = f(y) \quad (1)$$

where y is an n -composition vector; $f(y)$ is an arbitrary n -function vector, which does not contain t explicitly. If the system can be exactly lumped by an $\hat{n} \times n$ real constant matrix M with rank \hat{n} ($\hat{n} \leq n$), then for

$$\hat{y} = My \quad (2)$$

the lumped system can be described as

$$d\hat{y}/dt = Mf(\bar{M}\hat{y}) \quad (3)$$

where the subspace \mathcal{M} spanned by the row vectors of M is a fixed invariant one to the transpose of the Jacobian matrix $J^T(y)$ of $f(y)$ for any value of $y \in \Omega$, and \bar{M} is one of the generalized inverses of M (Ben-Israel and Greville, 1974) satisfying

$$M\bar{M} = I_{\hat{n}} \quad (4)$$

If $J^T(y)$ does not have a fixed invariant subspace which has a given dimension \hat{n} or satisfies some desired restriction, then eq. (3) can still be used to describe the lumped system approximately. In this case, one needs to find a subspace \mathcal{M} which meets the requirements and is as nearly $J^T(y)$ -invariant as possible. This lumping matrix is the best one for the given dimension \hat{n} and under the required restriction. The accuracy may not be satisfactory if \hat{n} is too small. When Ω is the whole n -dimensional composition space and M has orthonormal rows, M^T is the best choice of \bar{M} for approximate lumping (Li and Rabitz, 1990). Considering this we will choose orthonormal rows for M and consequently $\bar{M} = M^T$.

* Author to whom correspondence should be addressed.

For a nonisothermal first-order reaction system the kinetic equations are the following:

$$dy/dt = K(T)y \quad (5)$$

where $K(T)$ is the rate constant matrix, which is a function of temperature T . According to eq. (3) the lumped system can be represented as

$$\begin{aligned} d\hat{y}/dt &= \hat{K}(T)\hat{y} \\ &= MK(T)M^T\hat{y}. \end{aligned} \quad (6)$$

For the constrained lumping problem the lumping matrix M can be represented as

$$M = \begin{pmatrix} M_G \\ M_D \end{pmatrix} \quad (7)$$

where M_G is given and also required to satisfy $M_G M_G^T = I_{\hat{n}-r}$; M_D will be determined and satisfy $M_D M_D^T = I_r$ (where r is the row number of M_D) as well. The direct approach to determine the constrained lumping schemes with different \hat{n} has been presented in our previous paper (Li and Rabitz, 1991). This approach is based on the concept of the minimal $J^T(y)$ -invariant subspace over $\text{Im } M_G^T$. Again following the previous work on exact lumping, $J^T(y)$ can be decomposed into a linear combination of appropriate constant matrices A_k ($k = 1, 2, \dots, m$), i.e.

$$J^T(y) = \sum_{k=1}^m a_k(y) A_k \quad (8)$$

where m is less than n^2 and the A_k s are viewed as a basis of $J^T(y)$. When Ω is the whole n -dimensional space, the minimal simultaneously all A_k -invariant subspace over $\text{Im } M_G^T$ is the minimal $J^T(y)$ -invariant one over $\text{Im } M_G^T$.

In order to understand the basic idea of the direct approach in the application of the nonisothermal first-order reaction system, we will briefly draw from our previous paper about the basis of this method. It is well known that the minimal invariant subspace \mathcal{H} for an $n \times n$ matrix A over a given subspace $\text{Im } B$ coincides with

$$\mathcal{H} = \sum_{j=0}^{\infty} \text{Im}(A^j B) = \sum_{j=0}^{s-1} \text{Im}(A^j B) \quad (9)$$

for every integer s greater than or equal to the rank or the degree of a minimal polynomial for A in particular, $\mathcal{H} = \sum_{j=0}^{n-1} \text{Im}(A^j B)$ (Gohberg *et al.*, 1986). We know that

$$\sum_{j=0}^{s-1} \text{Im}(A^j B) = \text{Im}(B A B \dots A^{s-1} B) \quad (10)$$

and the orthogonal decomposition of the n -dimensional real space \mathcal{R}^n is

$$\mathcal{R}^n = \text{Im}(B A B \dots A^{s-1} B) \oplus \text{Ker} \begin{pmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{s-1} \end{pmatrix}. \quad (11)$$

In order to determine $\text{Im}(B A B \dots A^{s-1} B)$ we can first determine the kernel by solving the following equation

$$\begin{pmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{s-1} \end{pmatrix} X = 0. \quad (12)$$

Suppose the dimension of $\text{Im } X$ is $n-l$. After the determination of X the matrix representation M^T of the smallest A -invariant subspace \mathcal{H} with dimension l over $\text{Im } B$ can be determined by solving the equation

$$X^T M^T = 0. \quad (13)$$

It is straightforward to determine the minimal simultaneously A_k ($k = 1, 2, \dots, m$)-invariant subspace \mathcal{H} over the subspace $\text{Im } B$. We only need to determine X first by solving the following equation:

$$\begin{bmatrix} B^T \\ B^T A_1^T \\ \vdots \\ B^T (A_1^T)^{s_1-1} \\ \vdots \\ B^T \\ B^T A_m^T \\ \vdots \\ B^T (A_m^T)^{s_m-1} \end{bmatrix} X = 0 \quad (14)$$

where s_k ($k = 1, \dots, m$) is greater than or equal to the rank of A_k , and then solve eq. (13) to determine M . In the current problem $B = M_G^T$, $\mathcal{R}^n = Y_n$ and the resultant M is the exact lumping matrix containing M_G with the smallest row number l .

When we want to proceed further to find good-quality approximate lumping matrices with \hat{n} less than l , we need first to determine higher-dimensional $\text{Im } X$ which are as nearly as possible orthogonal to

$$\begin{bmatrix} M_G \\ M_G A_1^T \\ \vdots \\ M_G (A_1^T)^{s_1-1} \\ \vdots \\ M_G \\ M_G A_m^T \\ \vdots \\ M_G (A_m^T)^{s_m-1} \end{bmatrix}. \quad (15)$$

Then the resultant \mathcal{H} s will be as nearly all A_k -

invariant as possible. The corresponding M s are good approximate lumping matrices containing M_G with \hat{n} less than l . This consideration is equivalent to finding the subspace $\text{Im } X$, which is simultaneously as nearly orthogonal to $\text{Im } M_G^T$, $\text{Im}(M_G A_1^T)^T$, ..., $\text{Im}[M_G(A_1^T)^{s_1-1}]^T$, $\text{Im } M_G^T$, $\text{Im}(M_G A_2^T)^T$, ..., $\text{Im}[M_G(A_2^T)^{s_2-1}]^T$ as possible. This X can be readily determined by using the concept of the degree of coincidence between two subspaces given in our previous paper (Li and Rabitz, 1990).

Let $Q(G)_{(ki)}^T$ ($k = 1, 2, \dots, m; i = 0, 1, \dots, s_k - 1$) be the orthonormal matrix representation of $\text{Im}[M_G(A_k^T)^i]^T$. Using the Schmidt orthogonalization method one can transform $[M_G(A_k^T)^i]^T$ to $Q(G)_{(ki)}^T$. First we define a matrix

$$Y = \sum_{k=1}^m \sum_{i=0}^{s_k-1} Q(G)_{(ki)}^T Q(G)_{(ki)}. \quad (16)$$

If we choose an orthonormal basis for $\text{Im } X$, i.e.

$$X^T X = I_{n-\hat{n}}, \quad (17)$$

then the problem becomes the determination of X , which gives the smallest trace

$$\min_{X^T X = I_{n-\hat{n}}} \text{tr } X^T Y X. \quad (18)$$

The solution can be readily obtained by determining the eigenvalues and eigenvectors of Y (Bellman, 1970). The $n - \hat{n}$ eigenvectors with the smallest sum of their eigenvalues are X and the rest of the eigenvectors compose M^T . When all the eigenvalues are distinct, the solution for M with a specified \hat{n} is unique. If there exist multiple eigenvalues, the sets of eigenvectors with the same sum of eigenvalues are all solutions. When the eigenvectors of Y are arranged according to the nonincreasing order of their eigenvalues, the last $n - \hat{n}$ eigenvectors are X and the first \hat{n} eigenvectors are M^T . Therefore, the eigenvector matrix R of Y supplies all the best approximate lumping matrices with different \hat{n} .

There are two further issues we need to consider. First, sometimes $M_G A_i^T$ is a null matrix. In this case the contribution of A_i to the determination of the lumping matrix can be neglected. In order to avoid this situation, we can use the resultant M from other A_k with row number 1 higher than M_G as a new M_G to calculate $M_G A_i^T$. If $M_G A_i^T$ for the new M_G is still a null matrix, we can use the resultant M with row number 2 higher than the original M_G as a new M_G to calculate $M_G A_i^T$ and so on. Second, in order to satisfactorily assure that the resultant M_D is orthogonal to M_G , one can multiply M_G in eq. (15) by a large positive constant c .

For the nonisothermal first-order reaction system we have

$$J^T(y) = K^T(T). \quad (19)$$

Since the rate constant is an exponential function of temperature T , it is not easy to determine the basis matrices of $K^T(T)$. However, all the A_k s can be replaced by a set of rate constant matrices correspond-

ing to different temperatures in the desired temperature region. When the number of the rate constant matrices is large enough (i.e. some of these constant matrices compose a basis) and the temperature region is covered properly by the chosen temperature points (i.e. the different regions of temperature are appropriately weighted), the results should be the same or close to those obtained by using the basis matrices. Since this is easy to realize, the approach above is very useful for those systems whose Jacobian matrix cannot readily be decomposed to a linear combination of constant matrices. Let $K(T_i)$ be the rate constant matrix at temperature T_i , then eq. (15) becomes

$$\begin{bmatrix} M_G \\ M_G K(T_1) \\ \vdots \\ M_G K(T_1)^{s_1-1} \\ \vdots \\ M_G \\ M_G K(T_m) \\ \vdots \\ M_G K(T_m)^{s_m-1} \end{bmatrix}. \quad (20)$$

Thus the constrained lumping matrices with different \hat{n} can be obtained by the corresponding eigenvectors of Y .

If the subspace \mathcal{M} spanned by the row vectors of M is $J^T(y)$ -invariant, we have

$$MJ(y) = Q(y)M \quad (21)$$

where $Q(y)$ is an $\hat{n} \times \hat{n}$ matrix. It is easy to demonstrate that \mathcal{M} is also invariant to any analytic function of $J^T(y)$. Let $f^T[J^T(y)]$ be an analytic function of $J^T(y)$. It can be expanded in a Taylor series:

$$f^T[J^T(y)] = \sum_{i=0}^{\infty} c_i [J^T(y)]^i \quad (22)$$

where $[J^T(y)]^0 = I_n$ and c_i s are coefficients. It is easy to find that

$$f[J^T(y)] = \sum_{i=0}^{\infty} c_i [J(y)]^i. \quad (23)$$

Then we have

$$\begin{aligned} Mf[J^T(y)] &= M \sum_{i=0}^{\infty} c_i [J(y)]^i \\ &= \sum_{i=0}^{\infty} c_i [Q(y)]^i M \\ &= \hat{f}[Q^T(y)]M \end{aligned} \quad (24)$$

where

$$\hat{f}[Q^T(y)] = \sum_{i=0}^{\infty} c_i [Q^T(y)]^i \quad (25)$$

and we have used the relation of eq. (21) in the

deduction of eq. (24). Equation (24) shows that \mathcal{H} is $f^T[J^T(y)]$ -invariant. This is very useful for the first-order reaction system, because the analytic function $e^{K(T)\tau}$ of $K(T)$ can often be determined experimentally. The solution of eq. (5) is $e^{K(T)\tau}y(0)$. Let $y_1(0), y_2(0), \dots, y_n(0)$ be n linearly independent initial values of y and compose the matrix $Y(0)$. $y_1(\tau), y_2(\tau), \dots, y_n(\tau)$ are the corresponding solutions for $t = \tau$ and compose the matrix $Y(\tau)$. Then we have

$$Y(\tau) = e^{K(T)\tau} Y(0). \quad (26)$$

Since $Y(0)$ and $Y(\tau)$ can be determined experimentally and $Y(0)$ is nonsingular, $e^{K(T)\tau}$ will be obtained by

$$e^{K(T)\tau} = Y(\tau) Y^{-1}(0). \quad (27)$$

In many realistic problems, the rate constant matrix $K(T)$ is usually unknown in advance. Therefore, taking advantage of this situation we can use $e^{K(T)\tau}$ in eq. (20) instead of $K(T)$ to determine the constrained lumping matrices with different \hat{n} . Let $G(T_i) = e^{K(T_i)\tau}$. Then we have

$$\begin{bmatrix} M_G \\ M_G G(T_1) \\ \vdots \\ M_G G(T_1)^{s_1-1} \\ \vdots \\ M_G \\ M_G G(T_m) \\ \vdots \\ M_G G(T_m)^{s_m-1} \end{bmatrix}. \quad (28)$$

This approach will be illustrated by the Mobil "10-lump cracking model". The best constrained further lumped systems with $\hat{n} = 3-6$ valid in a given temperature region will be given.

3. THE MOBILE "10-LUMP CRACKING MODEL"

The method proposed above will be illustrated by the Mobil "10-lump model" of catalytic cracking pro-

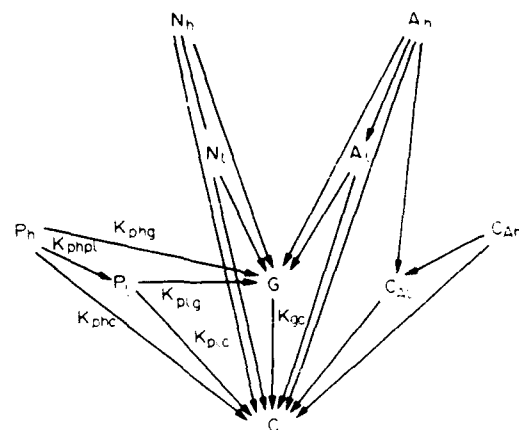


Fig. 1. 10-Lump cracking model kinetic scheme: P_l = wt % paraffinic molecules (mass spectroscopy analysis), 430–650°F; N_l = wt % naphthenic molecules (mass spectroscopy analysis), 430–650°F; C_{Al} = wt % carbon atoms among aromatic rings (n-d-M method), 430–650°F; A_l = wt % aromatic substituent groups, 430–650°F; P_h = wt % paraffinic molecules (mass spectroscopy analysis), 650+°F; N_h = wt % naphthenic molecules (mass spectroscopy analysis), 650+°F; C_{Ah} = wt % carbon atoms among aromatic rings (n-d-M method), 650+°F; A_h = wt % aromatic substituent groups, 650+°F; G = G-lump (C_3 , 430°F); C = C-lump (C_1 – C_4 + coke); $C_{Al} + P_l + N_l + A_l$ = LFO (430–650°F); $C_{Ah} + P_h + N_h + A_h$ = HFO (650+°F).

cess (Weekman, 1979; Jacob *et al.*, 1976). The scheme of this model is shown in Fig. 1. The composition vector is

$$y = (P_h N_h A_h C_{Ah} P_l N_l A_l C_{Al} G C)^T.$$

The corresponding rate constant matrix $K(T)$ is given in Fig. 2. The sum of P_h, N_h, A_h and C_{Ah} is called the heavy fuel oil (HFO) and the sum of P_l, N_l, A_l and C_{Al} is called the light fuel oil (LFO). The data of $K(T)$ for $T = 900^\circ\text{F}$ and the activation energies derived from temperatures of 900, 950 and 1000°F are available (Gross *et al.*, 1976). Using these data and weight % units for the concentration of the species, we obtain the $K(T)$ for $T = 900, 950$ and 1000°F as follows (in units of 10^3 h^{-1}):

$$K(900) = \begin{bmatrix} -83.55 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -122.07 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -166.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -20.49 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 20.70 & 0.00 & 0.00 & 0.00 & -33.29 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 22.50 & 0.00 & 0.00 & 0.00 & -74.33 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 19.00 & 0.00 & 0.00 & 0.00 & -22.13 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 50.00 & 5.86 & 0.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 \\ 55.00 & 84.70 & 63.00 & 0.00 & 23.85 & 66.15 & 18.50 & 0.00 & -4.40 & 0.00 \\ 7.85 & 14.87 & 34.20 & 14.63 & 9.44 & 8.18 & 3.63 & 1.00 & 4.40 & 0.00 \end{bmatrix}$$

$$K(950) = \begin{bmatrix} -83.86 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -122.51 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -167.38 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -20.67 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 20.80 & 0.00 & 0.00 & 0.00 & -33.42 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 22.60 & 0.00 & 0.00 & 0.00 & -74.58 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 19.09 & 0.00 & 0.00 & 0.00 & -22.32 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 50.23 & 5.89 & 0.00 & 0.00 & 0.00 & -1.01 & 0.00 & 0.00 \\ 55.17 & 84.97 & 63.52 & 0.00 & 23.93 & 66.36 & 18.65 & 0.00 & -4.45 & 0.00 \\ 7.89 & 14.94 & 34.54 & 14.78 & 9.49 & 8.22 & 3.67 & 1.01 & 4.45 & 0.00 \end{bmatrix}$$

$$K(1000) = \begin{bmatrix} -84.15 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -122.93 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -168.51 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -20.83 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 20.89 & 0.00 & 0.00 & 0.00 & -33.53 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 22.70 & 0.00 & 0.00 & 0.00 & -74.81 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 19.17 & 0.00 & 0.00 & 0.00 & -22.50 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 50.45 & 5.91 & 0.00 & 0.00 & 0.00 & -1.02 & 0.00 & 0.00 \\ 55.34 & 85.22 & 64.02 & 0.00 & 24.00 & 66.55 & 18.80 & 0.00 & -4.50 & 0.00 \\ 7.92 & 15.01 & 34.87 & 14.92 & 9.53 & 8.26 & 3.70 & 1.02 & 4.50 & 0.00 \end{bmatrix}$$

The $G(T_i) = e^{K(T_i)\tau}$ were computed with $\tau = 10^{-5}$ which was chosen because the significant dynamics occurred within 10τ :

$$G(900) = \begin{bmatrix} 0.4337 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2950 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1898 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.8147 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1166 & 0.0000 & 0.0000 & 0.0000 & 0.7168 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0851 & 0.0000 & 0.0000 & 0.0000 & 0.4755 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0807 & 0.0000 & 0.0000 & 0.0000 & 0.8015 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2422 & 0.0527 & 0.0000 & 0.0000 & 0.0000 & 0.9900 & 0.0000 & 0.0000 \\ 0.3803 & 0.5157 & 0.3085 & 0.0000 & 0.1982 & 0.4554 & 0.1622 & 0.0000 & 0.9570 & 0.0000 \\ 0.0694 & 0.1042 & 0.1788 & 0.1326 & 0.0849 & 0.0691 & 0.0363 & 0.0100 & 0.0430 & 1.0000 \end{bmatrix}$$

$$G(950) = \begin{bmatrix} 0.4323 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2937 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1875 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.8133 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1169 & 0.0000 & 0.0000 & 0.0000 & 0.7159 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0852 & 0.0000 & 0.0000 & 0.0000 & 0.4744 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0806 & 0.0000 & 0.0000 & 0.0000 & 0.8000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2423 & 0.0529 & 0.0000 & 0.0000 & 0.0000 & 0.9900 & 0.0000 & 0.0000 \\ 0.3810 & 0.5164 & 0.3097 & 0.0000 & 0.1987 & 0.4562 & 0.1634 & 0.0000 & 0.9565 & 0.0000 \\ 0.0698 & 0.1047 & 0.1799 & 0.1338 & 0.0854 & 0.0694 & 0.0367 & 0.0100 & 0.0435 & 1.0000 \end{bmatrix}$$

$$G(1000) = \begin{bmatrix} 0.4311 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2925 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1854 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.8120 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.1172 & 0.0000 & 0.0000 & 0.0000 & 0.7151 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0853 & 0.0000 & 0.0000 & 0.0000 & 0.4733 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0805 & 0.0000 & 0.0000 & 0.0000 & 0.7985 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2423 & 0.0531 & 0.0000 & 0.0000 & 0.0000 & 0.9899 & 0.0000 & 0.0000 \\ 0.3816 & 0.5171 & 0.3108 & 0.0000 & 0.1991 & 0.4569 & 0.1645 & 0.0000 & 0.9560 & 0.0000 \\ 0.0701 & 0.1051 & 0.1810 & 0.1350 & 0.0857 & 0.0698 & 0.0370 & 0.0101 & 0.0440 & 1.0000 \end{bmatrix}$$

	P_h	N_h	A_h	C_{ah}	P_L	N_L	A_L	C_{AL}	G	C
P_h	$-(K_{unp} + K_{phg} + K_{gnc})$	0	0	0	0	0	0	0	0	0
N_h	0	$-(K_{annl} + K_{nng} + K_{nnc})$	0	0	0	0	0	0	0	0
A_h	0	0	$-(K_{ghal} + K_{ang} + K_{anc} + K_{ahcal})$	0	0	0	0	0	0	0
C_{ah}	0	0	0	$-(K_{ahcal} + K_{cgh})$	0	0	0	0	0	0
P_L	$\nu_{hl} K_{ghal}$	0	0	0	$-(K_{plg} + K_{plc})$	0	0	0	0	0
N_L	0	$\nu_{hl} K_{annl}$	0	0	0	$-(K_{nlg} + K_{nnc})$	0	0	0	0
A_L	0	0	$\nu_{hl} K_{ghal}$	0	0	0	$-(K_{allg} + K_{gllc})$	0	0	0
C_{AL}	0	0	$\nu_{hl} K_{ahcal}$	$\nu_{hl} K_{ahcal}$	0	0	0	$-K_{lalc}$	0	0
G	$\nu_{hg} K_{unp}$	$\nu_{hg} K_{ang}$	$\nu_{hg} K_{ghg}$	0	$\nu_{lg} K_{plg}$	$\nu_{lg} K_{nlg}$	$\nu_{lg} K_{allg}$	0	$-K_{glc}$	0
C	$\nu_{hc} K_{ghc}$	$\nu_{hc} K_{nnc}$	$\nu_{hc} K_{anc}$	$\nu_{hc} K_{cgh}$	$\nu_{lc} K_{plc}$	$\nu_{lc} K_{nnc}$	$\nu_{lc} K_{allc}$	$\nu_{lc} K_{lalc}$	$\nu_{gc} K_{glc}$	0

Fig. 2. Rate constant matrix $K(T)$ of the 10-lump cracking model: ν_M = stoichiometric coefficient (mol. wt of HFO/mol. wt of LFO), ν_W = stoichiometric coefficient (mol. wt of HFO/mol. wt of gasoline), ν_M = stoichiometric coefficient (mol. wt of HFO/mol. wt of C-lump), ν_W = stoichiometric coefficient (mol. wt of LFO/mol. wt of gasoline), ν_L = stoichiometric coefficient (mol. wt of LFO/mol. wt of C-lump), ν_W = stoichiometric coefficient (mol. wt of gasoline/mol. wt of C-lump).

The goal of the catalytic cracking process is the production of gasoline. The C-lump (H_2 , H_2S , C_1 , C_4 and coke) is the undesired by-product. These two species correspond to y_9 and y_{10} of y . Therefore, we keep them unlumped and lump the other species to simplify this system. Hence the given part of the lumping matrix M is

$$M_G = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This information will be used in the following sections.

$$Y(K) = \begin{bmatrix} 0.71 & 1.80 & -0.11 & -0.14 & 0.07 & 0.78 & 0.09 & -0.01 & -0.05 & 0.00 \\ 1.80 & 7.33 & 0.27 & -0.16 & 0.14 & 1.78 & 0.14 & -0.01 & -0.06 & 0.00 \\ -0.11 & 0.27 & 8.67 & 0.46 & 0.21 & -0.11 & 0.02 & 0.03 & 0.13 & 0.00 \\ 0.14 & -0.16 & 0.46 & 0.30 & 0.07 & -0.19 & -0.02 & 0.02 & 0.11 & 0.00 \\ 0.07 & 0.14 & 0.21 & 0.07 & 0.05 & 0.08 & 0.02 & 0.01 & 0.02 & 0.00 \\ 0.78 & 1.78 & -0.11 & -0.19 & 0.08 & 0.87 & 0.11 & -0.01 & -0.07 & 0.00 \\ 0.09 & 0.14 & 0.02 & -0.02 & 0.02 & 0.11 & 0.02 & 0.00 & -0.01 & 0.00 \\ -0.01 & -0.01 & 0.03 & 0.02 & 0.01 & -0.01 & 0.00 & 0.00 & 0.01 & 0.00 \\ -0.05 & -0.06 & 0.13 & 0.11 & 0.02 & -0.07 & -0.01 & 0.01 & 10^4 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 10^4 \end{bmatrix}$$

3.4. The lumping schemes in the isothermal regime

In order to find the difference between $K(T)$ and $e^{K(T)\tau}$ in the determination of constrained lumping

schemes we first determine the constrained lumping schemes at 900 F by using $K(900)$ and $G(900)$, respectively. In this case eqs (20) and (28) become

$$R(K) = \begin{bmatrix} \lambda_1 = 10^4 & 10^4 & 8.7822 & 8.2302 & 0.6857 & 0.2433 & 0.0167 & 0.0007 & 0.0001 & 0.0000 \\ 0 & 0 & 0.0695 & -0.2395 & 0.5151 & -0.2174 & 0.2937 & -0.7059 & 0.1875 & 0.0739 \\ 0 & 0 & 0.3426 & -0.8695 & -0.3395 & 0.0887 & -0.0504 & 0.0302 & -0.0083 & -0.0023 \\ 0 & 0 & 0.9329 & 0.3537 & 0.0463 & 0.0485 & 0.0102 & -0.0052 & -0.0006 & 0.0001 \\ 0 & 0 & 0.0411 & 0.0484 & -0.3379 & -0.8708 & 0.2554 & 0.0531 & -0.1717 & -0.1613 \\ 0 & 0 & 0.0292 & -0.0105 & 0.0481 & -0.3758 & -0.6799 & 0.0522 & 0.5602 & 0.2768 \\ 0 & 0 & 0.0699 & -0.2424 & 0.6969 & -0.1852 & 0.0365 & 0.6173 & -0.1684 & -0.0750 \\ 0 & 0 & 0.0096 & -0.0197 & 0.1210 & -0.0699 & -0.6176 & -0.3377 & -0.6279 & -0.3004 \\ 0 & 0 & 0.0024 & 0.0025 & -0.0203 & -0.0616 & 0.0277 & -0.0099 & -0.4460 & 0.8922 \\ 1 & 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 1 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

schemes we first determine the constrained lumping schemes at 900 F by using $K(900)$ and $G(900)$, respectively. In this case eqs (20) and (28) become

$$\begin{bmatrix} M_G \\ M_G K(900) \\ \vdots \\ M_G K(900)^9 \end{bmatrix} \quad (29)$$

and

$$\begin{bmatrix} M_G \\ M_G G(900) \\ \vdots \\ M_G G(900)^9 \end{bmatrix} \quad (30)$$

Here we choose $s = n = 10$. Then using eq. (16) the symmetric matrices Y and their eigenvector matrices R are determined. In order to force M_G^T to be located on the first two columns of R and the lumped species to be composed of the other eight original species (correspondingly the last two elements of each column of M_D are zero), M_G in the first row of eqs (29) and (30) are multiplied by 100. Let $Y(K)$, $Y(G)$ and $R(K)$, $R(G)$ represent the corresponding symmetric matrices and their eigenvector matrices for using $K(900)$ and $G(900)$, respectively. The eigenvalues are also given right above the corresponding eigenvector matrix.

In the case of using $K(900)$ the resultant $Y(K)$ is the following:

The eigenvalues λ_i of $Y(K)$ arranged in nonincreasing order and the eigenvector matrix $R(K)$, whose eigenvectors are arranged according to the order of their eigenvalues, are given below:

According to the direct approach the first three columns on the left of $R(K)$ compose the best constrained approximate lumping matrix with $\hat{n} = 3$, the first four columns compose the best constrained approximate lumping matrix with $\hat{n} = 4$ and so on. Since the last three eigenvalues are equal to or almost equal to zero, the first seven columns of $R(Y)$ compose an almost exact lumping matrix. From eq. (6) we know that

$$\hat{K}(T) = M K(T) M^T \quad (31)$$

Then we have the rate constant matrix for the lumped system with $\hat{n} = 7$ at 900 F as follows:

$$\hat{K}(900) = \begin{bmatrix} -4.4000 & 0.0000 & 97.1113 & -81.1857 & 51.9774 & -23.8956 & -12.6995 \\ 4.4000 & 0.0000 & 39.0312 & -4.1559 & 2.2088 & -16.8466 & -2.6924 \\ 0.0000 & 0.0000 & -158.9405 & -17.2578 & 0.4065 & -7.8806 & 0.6869 \\ 0.0000 & 0.0000 & -17.9688 & -117.5942 & -13.7646 & -0.8830 & 0.0376 \\ 0.0000 & 0.0000 & 7.2540 & -28.9118 & -80.1459 & 18.3513 & -12.7016 \\ 0.0000 & 0.0000 & -13.9865 & 3.4880 & 14.3366 & -26.7782 & -0.8136 \\ 0.0000 & 0.0000 & -11.1369 & -1.1708 & -20.2488 & 3.8874 & -37.0395 \end{bmatrix}$$

When we use the first \hat{n} ($\hat{n} < 7$) columns of $R(K)$ to compose the lumping matrix, the resultant lumped rate constant matrix is the $\hat{n} \times \hat{n}$ submatrix in the top left-hand corner of the above matrix. Therefore, this matrix supplies all $\hat{K}(900)$ for $\hat{n} = 3-7$.

For the initial composition ($y_1 = y_6 = \frac{1}{6}$, others are zero) we obtained the evolutions of the concentration of y_6 by solving eqs (5) and (6) (for $\hat{n} = 4-7$). The results are shown in Fig. 3. One can see that, when \hat{n} becomes larger, the solution of the lumped system is closer to that of the original one. For $\hat{n} = 7$ the lumping is almost exact.

Following the same procedure we use $G(900)$ instead of $K(900)$ to determine the constrained lumping matrices for different \hat{n} . The resultant $Y(G)$ is the following:

$$Y(G) = \begin{bmatrix} 1.30 & 1.44 & 0.77 & 0.02 & 0.98 & 1.46 & 0.97 & 0.00 & 1.74 & 0.05 \\ 1.44 & 1.61 & 0.86 & 0.03 & 1.08 & 1.62 & 1.05 & 0.00 & 1.97 & 0.09 \\ 0.77 & 0.86 & 0.64 & 0.49 & 0.68 & 0.82 & 0.54 & 0.06 & 0.93 & 1.16 \\ 0.02 & 0.03 & 0.49 & 1.40 & 0.31 & -0.11 & -0.06 & 0.16 & -0.31 & 2.96 \\ 0.98 & 1.08 & 0.68 & 0.31 & 0.82 & 1.07 & 0.74 & 0.04 & 1.20 & 0.72 \\ 1.46 & 1.62 & 0.82 & -0.11 & 1.07 & 1.64 & 1.08 & -0.01 & 2.00 & -0.21 \\ 0.97 & 1.05 & 0.54 & -0.06 & 0.74 & 1.08 & 0.75 & -0.01 & 1.25 & -0.08 \\ 0.00 & 0.00 & 0.06 & 0.16 & 0.04 & -0.01 & -0.01 & 0.02 & -0.03 & 0.33 \\ 1.74 & 1.97 & 0.93 & -0.31 & 1.20 & 2.00 & 1.25 & -0.03 & 10^4 & -0.75 \\ 0.05 & 0.09 & 1.16 & 2.96 & 0.72 & -0.21 & -0.08 & 0.33 & -0.75 & 10^4 \end{bmatrix}$$

The eigenvalues λ_i of $Y(G)$ arranged in nonincreasing order and the eigenvector matrix $R(G)$ arranged according to the order of their eigenvalues are given below:

$$\lambda_i = 10^4, \quad 10^4, \quad 6.4681, \quad 1.6352, \quad 0.0642, \quad 0.0142, \quad 0.0005, \quad 0.0002, \quad 0.0000, \quad 0.0000$$

$$R(G) = \begin{bmatrix} 0 & 0 & 0.4475 & -0.0623 & -0.0095 & -0.2227 & 0.0646 & 0.0160 & 0.0217 & -0.8610 \\ 0 & 0 & 0.4953 & -0.0625 & -0.4301 & -0.0774 & -0.3854 & -0.4522 & -0.3865 & 0.2396 \\ 0 & 0 & 0.2769 & 0.2811 & -0.2664 & 0.8340 & -0.0097 & 0.2385 & 0.1183 & -0.0825 \\ 0 & 0 & 0.0381 & 0.9203 & 0.0128 & -0.3109 & -0.0871 & -0.1119 & 0.1840 & 0.0295 \\ 0 & 0 & 0.3455 & 0.1508 & 0.4648 & 0.1052 & 0.5944 & -0.1031 & -0.4889 & 0.1667 \\ 0 & 0 & 0.4976 & -0.1526 & -0.2024 & -0.3090 & 0.2945 & 0.3770 & 0.4580 & 0.3925 \\ 0 & 0 & 0.3307 & -0.0928 & 0.6976 & 0.1121 & -0.5522 & -0.0005 & 0.2576 & 0.1069 \\ 0 & 0 & 0.0042 & 0.1080 & -0.0063 & -0.1820 & -0.3075 & 0.7571 & -0.5358 & 0.0190 \\ 1 & 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 1 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

Observing the eigenvalues of $R(G)$ we found that the last four eigenvalues are equal to or almost equal to zero. Therefore, the first six columns of $R(G)$ compose an almost exact lumping matrix. Using eq. (31) the resultant rate constant matrix for the lumped system with $\hat{n} = 6$ at 900°F is the following:

$$\hat{K}(900) = \begin{bmatrix} -4.4000 & 0.0000 & 131.2834 & 0.7743 & -43.1329 & 17.8802 \\ 4.4000 & 0.0000 & 29.4419 & 21.6056 & -10.1357 & 19.7656 \\ 0.0000 & 0.0000 & -73.7046 & -2.2557 & 29.0315 & -12.7847 \\ 0.0000 & 0.0000 & -2.2309 & -32.3531 & 6.1642 & -36.2268 \\ 0.0000 & 0.0000 & 41.2758 & 8.9664 & -56.9739 & 33.7715 \\ 0.0000 & 0.0000 & -20.1729 & -41.2757 & 29.5716 & -135.6620 \end{bmatrix}$$

Similarly this matrix supplies all $\hat{K}(900)$ with $\hat{n} < 6$ by the $\hat{n} \times \hat{n}$ submatrices in the top left-hand corner of the above matrix. The comparison of y_6 between the exact solution and the solutions given by the lumped models with $\hat{n} = 3-6$ is shown in Fig. 4. When $\hat{n} = 6$ the coincidence between the exact and the lumped models is very good.

From the results obtained by using $K(900)$ and $G(900)$ one can find that $G(T)$ gives the better results. The reason is not entirely clear. Possibly the lumping schemes given by $K(T)$ are valid in the whole n -dimensional space, while the lumping schemes obtained from $G(T)$ are suitable for the whole composi-

tion region (i.e. all y_i being nonnegative and $\sum_i y_i = 1$). Considering the results we will determine the lumping schemes validated in the temperature region 900-1000°F by using $G(T)$.

3B. The lumping scheme for the nonisothermal regime

Since $G(T)$ for different temperatures (900–1000°F) are very close to one another, it is enough only to choose three matrices $G(900)$, $G(950)$ and $G(1000)$ to determine the lumping schemes for this temperature region.

Utilizing eqs (16) and (28) and following the same procedure as that in Section 3A, we obtain the symmetric matrix $Y(G)$, its eigenvalues and eigenvector matrix $R(G)$:

$$Y(G) = \begin{bmatrix} 3.91 & 4.32 & 2.31 & 0.05 & 2.95 & 4.37 & 2.91 & 0.01 & 5.23 & 0.16 \\ 4.22 & 4.81 & 2.60 & 0.08 & 3.24 & 4.84 & 3.16 & 0.01 & 5.89 & 0.27 \\ 2.31 & 2.60 & 1.93 & 1.47 & 2.05 & 2.46 & 1.63 & 0.17 & 2.80 & 3.49 \\ 0.05 & 0.08 & 1.47 & 4.23 & 0.94 & -0.32 & -0.19 & 0.50 & -0.94 & 8.93 \\ 2.95 & 3.24 & 2.05 & 0.94 & 2.47 & 3.21 & 2.22 & 0.11 & 3.60 & 2.15 \\ 4.37 & 4.84 & 2.46 & -0.32 & 3.21 & 4.93 & 3.25 & -0.04 & 5.99 & -0.65 \\ 2.91 & 3.16 & 1.63 & -0.18 & 2.22 & 3.25 & 2.28 & -0.02 & 3.76 & -0.25 \\ 0.01 & 0.01 & 0.17 & 0.50 & 0.11 & -0.04 & -0.02 & 0.06 & -0.10 & 1.00 \\ 5.23 & 5.89 & 2.80 & -0.94 & 3.60 & 5.99 & 3.76 & -0.10 & 3 \times 10^4 & -2.25 \\ 0.16 & 0.27 & 3.49 & 8.93 & 2.15 & -0.65 & -0.25 & 1.00 & -2.25 & 3 \times 10^4 \end{bmatrix}$$

$$\lambda_i = 3 \times 10^4, \quad 3 \times 10^4, \quad 19.4200, \quad 4.9500, \quad 0.1932, \quad 0.0428, \quad 0.0014, \quad 0.0005, \quad 0.0000, \quad 0.0000$$

$$R(G) = \begin{bmatrix} 0 & 0 & 0.4472 & -0.0622 & -0.0102 & -0.2227 & 0.0633 & 0.0149 & 0.0110 & -0.8614 \\ 0 & 0 & 0.4948 & -0.0622 & -0.4302 & -0.0779 & -0.3829 & -0.4527 & -0.3850 & 0.2458 \\ 0 & 0 & 0.2771 & 0.2807 & -0.2669 & 0.8340 & -0.0107 & 0.2381 & 0.1172 & -0.0841 \\ 0 & 0 & 0.0383 & 0.9206 & 0.0138 & -0.3101 & -0.0862 & -0.1130 & 0.1839 & 0.0275 \\ 0 & 0 & 0.3456 & 0.1500 & 0.4645 & 0.1046 & 0.5952 & -0.0974 & -0.4879 & 0.1719 \\ 0 & 0 & 0.4973 & -0.1522 & -0.2033 & -0.3093 & 0.2937 & 0.3769 & 0.4643 & 0.3857 \\ 0 & 0 & 0.3319 & -0.0934 & 0.6975 & 0.1125 & -0.5514 & -0.0040 & 0.2594 & 0.1044 \\ 0 & 0 & 0.0042 & 0.1077 & -0.0064 & -0.1826 & -0.3121 & 0.7575 & -0.5321 & 0.0251 \\ 1 & 0 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0 & 1 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix}$$

Comparing the resultant $R(G)$ with that for isothermal condition in Section 3A, one can see that the eigenvector matrices $R(G)$ are almost the same. Since we use three $G(T_i)$, the eigenvalues should be nearly 3 times of the eigenvalues for the isothermal condition. This is found to be true. Therefore, the first six columns of $R(G)$ will supply an almost exact lumping matrix for 900–1000°F. Using eq. (31) the lumped rate constant matrices for 900, 950 and 1000°F are as follows:

$$\hat{K}(900) = \begin{bmatrix} -4.4000 & 0.0000 & 131.2420 & 0.7763 & -43.2836 & 17.8111 \\ 4.4000 & 0.0000 & 29.4447 & 21.5948 & -10.1565 & 19.7626 \\ 0.0000 & 0.0000 & -73.6407 & -2.2723 & 29.0700 & -12.7736 \\ 0.0000 & 0.0000 & -2.2644 & -32.3174 & 6.1810 & -36.1860 \\ 0.0000 & 0.0000 & 41.3131 & 8.9754 & -57.0402 & 33.7829 \\ 0.0000 & 0.0000 & -20.1893 & -41.2263 & 29.5970 & -135.6848 \end{bmatrix}$$

$$\hat{K}(950) = \begin{bmatrix} -4.4500 & 0.0000 & 131.7775 & 0.8610 & -43.4412 & 18.1462 \\ 4.4500 & 0.0000 & 29.6477 & 21.8202 & -10.2328 & 19.9808 \\ 0.0000 & 0.0000 & -73.9510 & -2.3267 & 29.1827 & -12.9555 \\ 0.0000 & 0.0000 & -2.3185 & -32.5687 & 6.2528 & -36.4184 \\ 0.0000 & 0.0000 & 41.4835 & 9.0604 & -57.3446 & 34.0363 \\ 0.0000 & 0.0000 & -20.4050 & -41.4819 & 29.8297 & -136.5938 \end{bmatrix}$$

$$\hat{K}(1000) = \begin{bmatrix} -4.5000 & 0.0000 & 132.2843 & 0.9428 & -43.5854 & 18.4713 \\ 4.5000 & 0.0000 & 29.8362 & 22.0337 & -10.3181 & 20.1938 \\ 0.0000 & 0.0000 & -74.2425 & -2.3792 & 29.2919 & -13.1329 \\ 0.0000 & 0.0000 & -2.3706 & -32.7989 & 6.3230 & -36.6427 \\ 0.0000 & 0.0000 & 41.6453 & 9.1425 & -57.6298 & 34.2775 \\ 0.0000 & 0.0000 & -20.6144 & -41.7281 & 30.0547 & -137.4624 \end{bmatrix}$$

Similarly these matrices supply all the $\hat{K}(900)$, $\hat{K}(950)$ and $\hat{K}(1000)$ with $\hat{n} < 6$ by the $\hat{n} \times \hat{n}$ submatrices in the top left-hand corner of the above matrices. The comparisons of y_9 and y_{10} between the exact solutions and the solutions given by the lumped models with $\hat{n} = 3-6$ and $T = 900, 950$ and 1000°F are shown in Figs 5–10. The initial compositions chosen by Coxson and Bishoff (1987) are adopted here: (a) paraffinic = (0.3, 0.1, 0.15, 0.15, 0.2, 0.05, 0.03, 0.02, 0, 0); (b)

aromatic = (0.1, 0.1, 0.2, 0.4, 0.05, 0.05, 0.05, 0.05, 0, 0); and (c) naphthenic = (0.15, 0.4, 0.1, 0.08, 0.07, 0.2, 0, 0, 0, 0). They represent the basic charge compositions. To save space we do not give all the results for different initial compositions and temperatures. They slightly differ in accuracy for $\hat{n} = 3$ or 4, but they have

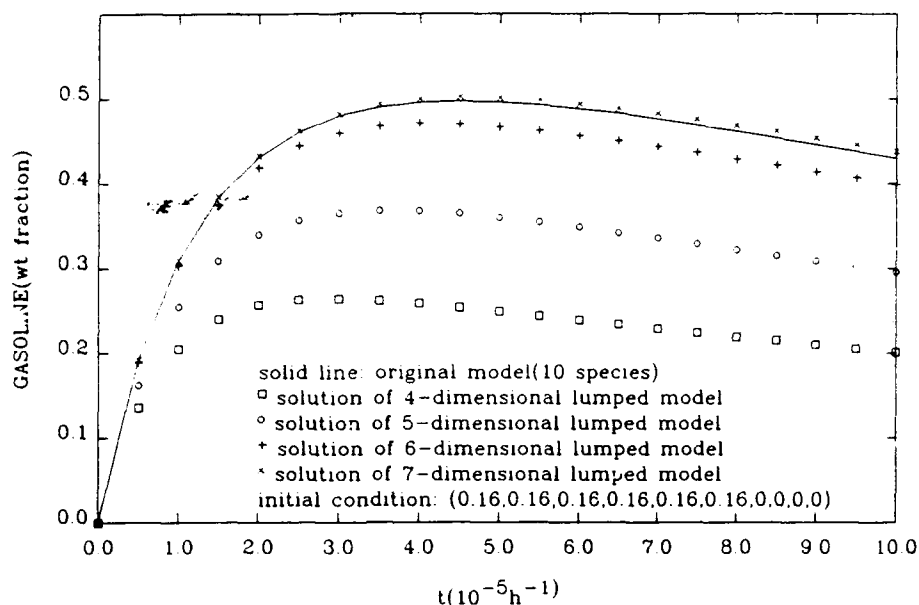


Fig. 3. Comparison of y_g (gasoline) for the original model and the isothermal lumped models obtained by using $K(900)$.

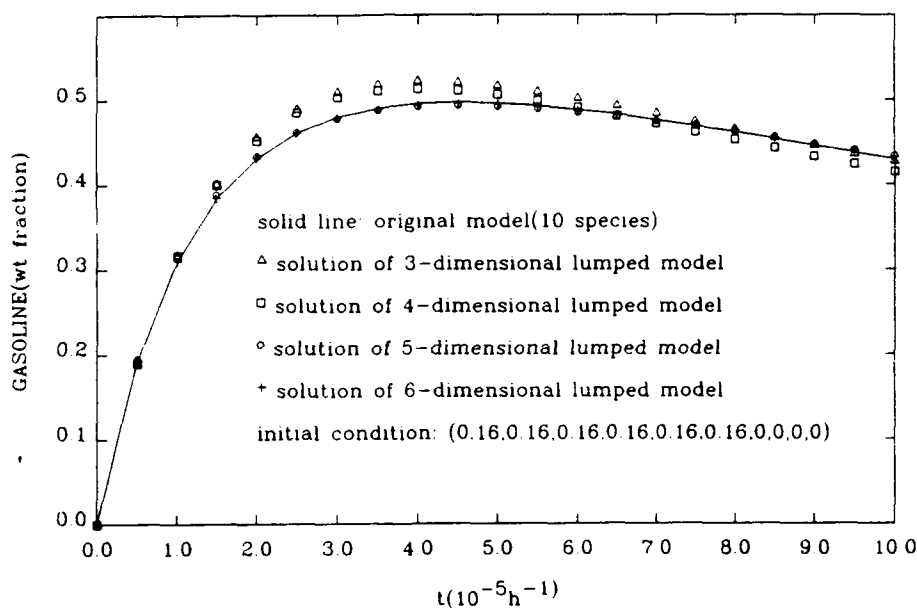


Fig. 4. Comparison of y_g (gasoline) for the original model and the lumped models obtained by using $G(900)$.

a similar accuracy for $\hat{n} = 5$ or 6. When $\hat{n} = 6$ the solutions for the lumped model in all these conditions are almost exactly the same as those of the original model. When $\hat{n} = 5$ the coincidence between the exact and the lumped models is very good. Considering that there exists experimental error in practice the lumped model with $\hat{n} = 5$ is adequate and the lumped model with $\hat{n} = 4$ is acceptable. Even if $\hat{n} = 3$, for most conditions the lumped model approximates the original system quite well. All these results show that the direct approach can be employed to determine the best constrained approximate lumping schemes for

the first-order reaction system under nonisothermal conditions.

4. CONCLUSION AND DISCUSSION

In the present paper, we have shown that the direct approach to determining the constrained approximate lumping schemes for an arbitrary reaction system can be employed to the determination of the lumping schemes for a first-order reaction system under the isothermal and nonisothermal conditions.

In the nonisothermal case the rate constant matrix $K(T)$ is a function of temperature T and the constant

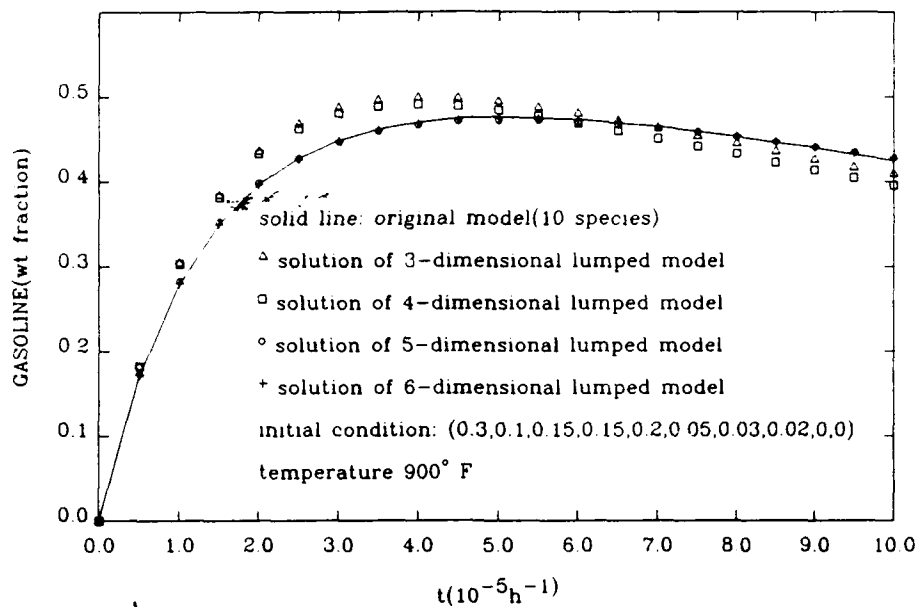


Fig. 5. Comparison of y_0 (gasoline) at $T = 900^\circ\text{F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

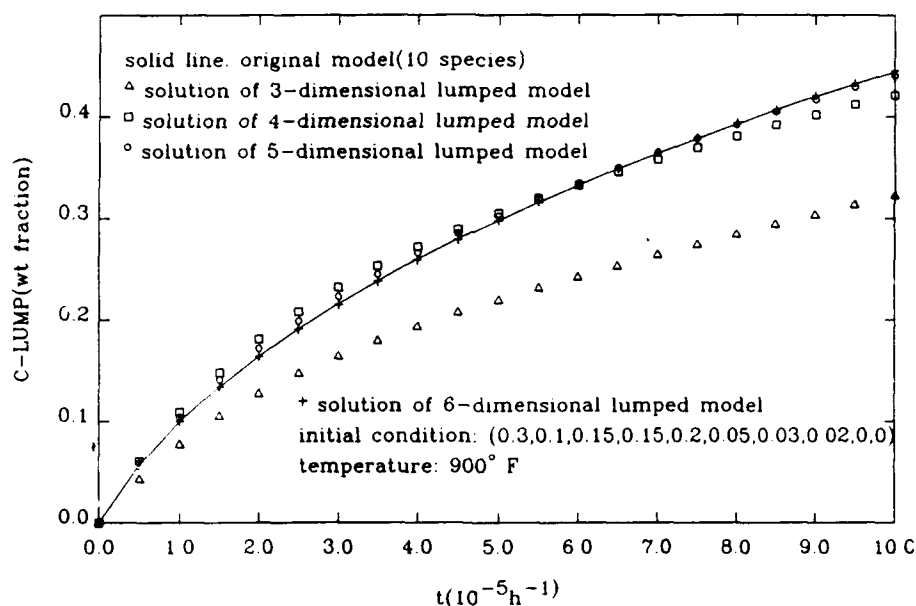


Fig. 6. Comparison of y_{10} (C-lump) at $T = 900^\circ\text{F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

basis matrices of $K(T)$ are not easy to determine. In this case one can use a set of $K(T_i)$ for different given temperatures, which properly cover the desired temperature region, instead of the basis matrices of $K(T)$.

If the subspace \mathcal{M} spanned by the row vectors of the lumping matrix is invariant to the transpose of the Jacobian matrix $J^T(y)$ of the kinetic equations, \mathcal{M} is also invariant to any analytic function of $J^T(y)$. For the first-order reaction system $J^T(y)$ is $K^T(T)$ and $[e^{K(T)\tau}]^T$ is an analytic function of $K^T(T)$. Therefore, one can use $e^{K(T)\tau}$ instead of $K(T)$ to determine the constrained approximate lumping matrices. The re-

sult of the present paper shows that the lumping schemes obtained by using $e^{K(T)\tau}$ are even better than those by using $K(T)$. Since $e^{K(T)\tau}$ can be determined experimentally, using $e^{K(T)\tau}$ is more advantageous.

The Mobil "10-lump cracking model" was used to illustrate this approach. The results show that this model can be adequately reduced to lumped ones with five or six additionally lumped species. The accuracy of the lumping schemes validated for the temperature range $T = 900-1000^\circ\text{F}$ is almost the same as that for $T = 900^\circ\text{F}$. This is because that the rate constants do not change much in this temperature range. For a

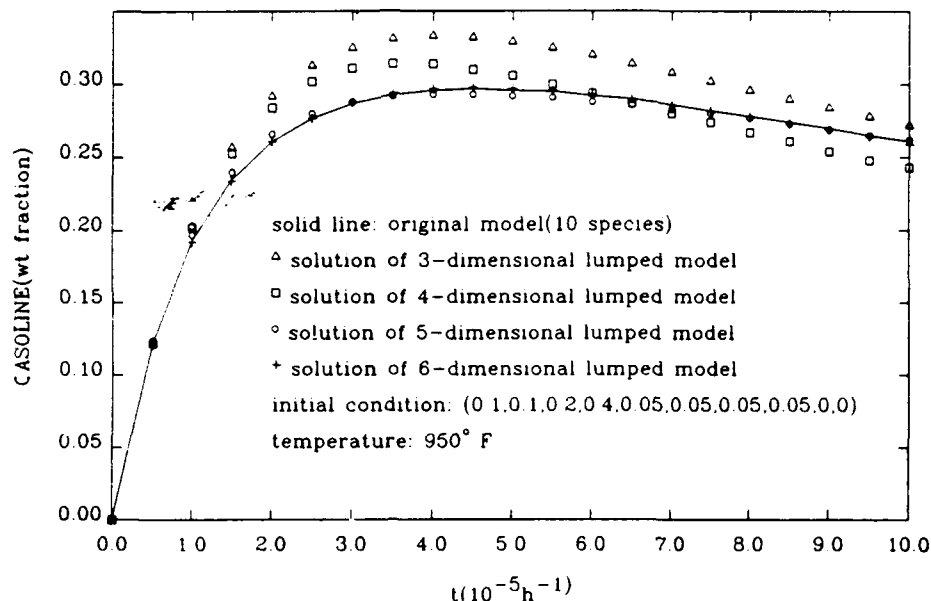


Fig. 7. Comparison of y_9 (gasoline) at $T = 950^\circ\text{F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

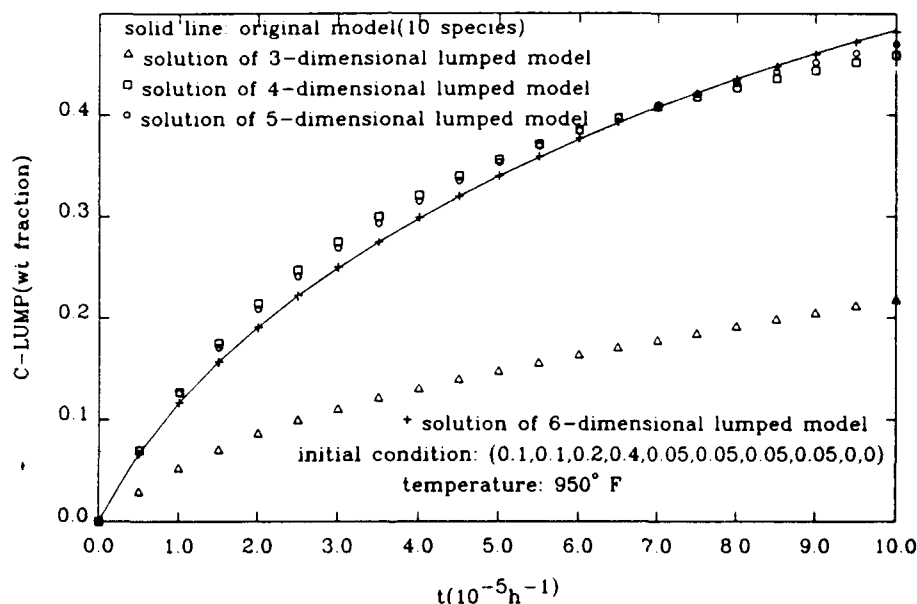


Fig. 8. Comparison of y_{10} (C-lump) at $T = 950^\circ\text{F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

wider range of temperature, the difference between the lumping schemes validated in the large temperature range and that for a given temperature in the same range will become larger.

The approach presented in this paper is not only applicable to first-order reaction systems but also to other ones under nonisothermal conditions. Let us consider the general case of a nonisothermal reaction system. It can be described as

$$\begin{aligned} dy/dt &= f(y, T) \\ dT/dt &= g(y, T). \end{aligned} \quad (32)$$

Let

$$\begin{aligned} z &= (y^T T)^T \\ h(y, T) &= [f^T(y, T) g(y, T)]^T. \end{aligned} \quad (33)$$

Then eq. (32) can be rewritten as

$$dz/dt = h(z). \quad (34)$$

The exact lumping of eq. (34) can be considered in the same way as that of eq. (1), except that the last "species" T is required unlumped (this means that the lumping matrix M must have a given row e_{n+1}).

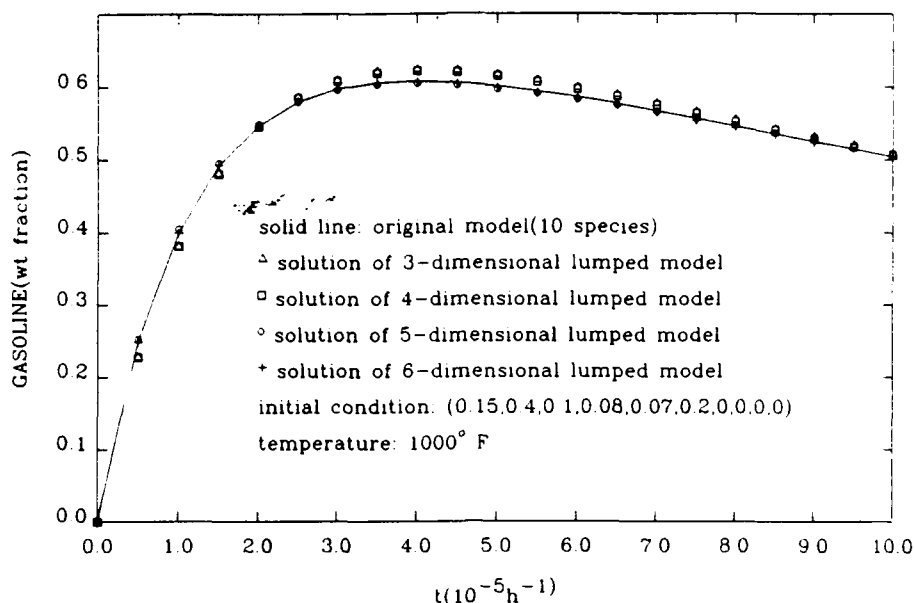


Fig. 9. Comparison of y_9 (gasoline) at $T = 1000^\circ \text{ F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

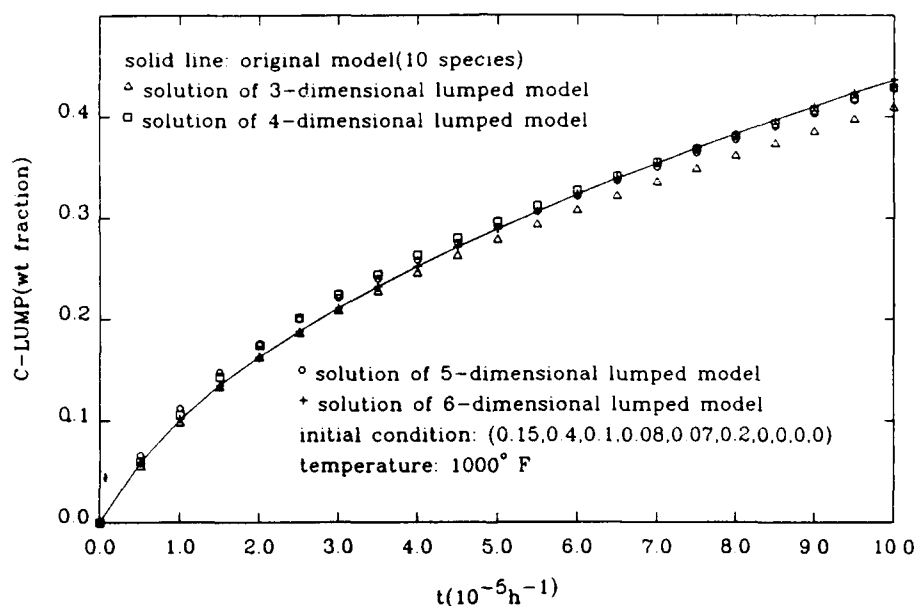


Fig. 10. Comparison of y_{10} (C-lump) at $T = 1000^\circ \text{ F}$ for the original model and the lumped models obtained by using $G(900)$, $G(950)$ and $G(1000)$.

Considering that the rate constants are exponential functions of temperature, the constant basis matrices of the transpose of the Jacobian matrix $J^T(\mathbf{z})$ of $\mathbf{h}(\mathbf{z})$ cannot generally be determined. However, for most reaction systems it may be decomposed as

$$J^T(\mathbf{z}) = \sum_{k=1}^m a_k(\mathbf{y}) A_k(T). \quad (35)$$

We need to find a fixed invariant subspace containing at least the unit vector \mathbf{e}_{n+1} simultaneously for all $A_k(T)$ in the desired region of T . If the constant basis matrices for every $A_k(T)$ are known, the fixed in-

variant subspaces of $J^T(\mathbf{z})$ are just the common fixed invariant subspaces to all these constant matrices. However, the $A_k(T)$ s are like $K(T)$ and their constant basis matrices are not easy to determine. Therefore, the approach to determine the fixed invariant subspaces of $K(T)$ presented in this paper can be employed to $A_k(T)$. We only need to properly choose a sufficient number of temperature T_i in the desired region and then to calculate the corresponding $A_k(T_i)$. Using equations similar to eqs (15) and (16) one can determine the constrained lumping matrices with different dimensions for any nonisothermal reaction sys-

tem. In order to obtain a good result the number of constant matrices for different temperatures may be quite large, but the computational effort is not very expensive, because the computation only contains matrix multiplication and determination of the eigenvalues and eigenvectors for a symmetric matrix. In conclusion, this approach is an easy way to determine constrained lumping schemes for any reaction system under nonisothermal conditions.

Acknowledgement—The authors acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

NOTATION

Scalars

$a_k(y)$	k th coefficient of the decomposition of $J^T(y)$
c	constant
c_i	coefficient
$g(y, T)$	derivative function of temperature
k	integer
l	integer
m	integer
\mathcal{M}	subspace spanned by the row vectors of M
n	dimension of vector y
\hat{n}	dimension of vector \hat{y}
r	row number of M_D
\mathcal{R}^n	n -dimensional real space
s	integer
s_k	rank of A_k
T	temperature
T_i	temperature
t	time
Y_n	n -dimensional composition space
y_k	k th element of vector y

Vectors and matrices

Capital letters represent matrices, bold-face lower-case letters represent vectors.

A	constant matrix
A_k	basis matrix of $J^T(y)$
B	constant matrix
e_{n+1}	unit vector with 1 as its $n+1$ entry, 0 for others
$f[J^T(y)]$	analytic function of $J^T(y)$
$\hat{f}[Q^T(y)]$	analytic function of $Q^T(y)$
$f(y)$	n -dimensional function vector
$\hat{f}(\hat{y})$	\hat{n} -dimensional function vector
$G(T)$	defined as $e^{K(T)\tau}$
$h(z)$	defined as $[f^T(z)g(z)]^T$
I	identity matrix
$J(y)$	Jacobian matrix of $f(y)$
$J(z)$	Jacobian matrix of $h(z)$
$K(T)$	rate constant matrix at temperature T

$\tilde{K}(T)$	rate constant matrix of the lumped system at temperature T
M	lumping matrix
M_D	determined submatrix of M
M_G	given submatrix of M
\bar{M}	generalized inverse of M satisfying $M\bar{M} = I_{\hat{n}}$
$Q(G)_{(ki)}^T$	matrix representation of $\text{Im}[M_G(A_k^T)^i]^T$ with orthonormal columns
$Q(y)$	$\hat{n} \times \hat{n}$ function matrix
$R(K)$	eigenvector matrix of $Y(K)$
$R(G)$	eigenvector matrix of $Y(G)$
X	$n \times (n - \hat{n})$ matrix
y	n -dimensional variable vector
\hat{y}	\hat{n} -dimensional variable vector
Y	symmetric matrix
$Y(K)$	symmetric matrix determined by $K(T)$
$Y(G)$	symmetric matrix determined by $G(T)$
$Y(0)$	defined as $[y_1(0) y_2(0) \dots y_n(0)]$
$Y(\tau)$	defined as $[y_1(\tau) y_2(\tau) \dots y_n(\tau)]$
z	defined as $(y^T T)^T$

Greek letters

λ_i	i th eigenvalue of matrix $Y(K)$ or $Y(G)$
τ	time
Ω	desired region of the composition space

Symbol

\wedge	any property related to the lumped system
----------	---

REFERENCES

- Bellman, R., 1970, *Introduction to Matrix Analysis*. McGraw-Hill, New York.
- Ben-Israel, A. and Greville, T. N. E., 1974, *Generalized Inverse: Theory and Applications*. John Wiley, New York.
- Coxson, P. G. and Bischoff, K. B., 1987, Lumping strategy. 1. Introductory techniques and applications of cluster analysis. *Ind. Engng Chem. Res.* **26**, 1239–1248.
- Gross, B., Jacob, S. M., Nace, D. M. and Voltz, S. E., 1976, Simulation of catalytic cracking process. US Patent 3,960,707.
- Gohberg, I., Lancaster, P. and Rodman, L., 1986, *Invariant Subspaces of Matrices with Applications*. John Wiley, New York.
- Jacob, S. M., Gross, B., Voltz, S. E. and Weekman, V. W., Jr., 1976, A lumping and reaction scheme for catalytic cracking. *A.I.Ch.E. J.* **22**, 701–713.
- Li, G. and Rabitz, H., 1989, A general analysis of exact lumping in chemical kinetics. *Chem. Engng Sci.* **44**, 1413–1430.
- Li, G. and Rabitz, H., 1990, A general analysis of approximate lumping in chemical kinetics. *Chem. Engng Sci.* **45**, 977–1002.
- Li, G. and Rabitz, H., 1991, New approaches to determination of constrained lumping schemes for a reaction system in the whole composition space. *Chem. Engng Sci.* **46**, 95–111.
- Weekman, V. W., Jr., 1979, Lumps, models, and kinetics in practice. *A.I.Ch.E. Monogr. Ser.* **75**(11), 3–29.

Appendix J

10. A General Lumping Analysis of a Reaction System Coupled with Diffusion,
G. Li and H. Rabitz, Chem. Eng. Sci., in press.

**A GENERAL LUMPING ANALYSIS OF A REACTION SYSTEM
COUPLED WITH DIFFUSION**

Genyuan Li and Herschel Rabitz*

Department of Chemistry

Princeton University

Princeton, New Jersey 08540

* Author to whom correspondence should be addressed.

Abstract

A general lumping analysis of a reaction system coupled with diffusion is presented. This analysis can be applied to any reaction system with n species for both steady-state and transient conditions. Here we consider lumping by means of an $\hat{n} \times n$ constant matrix M with rank \hat{n} ($\hat{n} \leq n$). When the diffusivity is independent of position and concentration vectors r and y , it is found that under steady-state conditions a reaction system having species concentration vector $y(r)$ coupled with diffusion is exactly lumpable if and only if there exist nontrivial fixed $J^T(y(r))D^{-1}$ -invariant subspaces \mathcal{M} (here $J^T(y(r))$ is the transpose of the Jacobian matrix for the chemical reaction rate vector $f(y(r))$ and D^{-1} is the inverse of the constant effective diffusivity matrix), no matter what value $y(r)$ takes; under transient conditions there exist simultaneously D - and $J^T(y(r,t))$ -invariant subspaces \mathcal{M} . When D is a function of position or concentrations, \mathcal{M} is simultaneously invariant to $J^T(y)$ and $D(r)$, $D(y(r))$ or $D(y(r,t))$. The same approach to determine the constrained approximate lumping schemes for a non-diffusion system can be used in a reaction-diffusion one except that the constant basis matrices A_k 's of $J^T(y)$ are replaced by $B_k = A_k D^{-1}$ under steady-state conditions or the extra matrix D is added under transient conditions. For nonconstant D the basis constant matrices D_i 's of $D(r)$, $D(y(r))$ or $D(y(r,t))$ are added.

I. INTRODUCTION

The general analyses of exact and approximate lumping in chemical kinetics have been presented in our previous papers(Li and Rabitz, 1989, 1990a, 1990b). In those papers we only considered homogeneous reaction systems without diffusion. In realistic problems many reaction systems are coupled with diffusion, which may modify greatly the behavior of the systems. Therefore, a general lumping analysis for reaction systems coupled with diffusion is necessary. When we consider a reaction system coupled with diffusion, we need to study both steady-state and transient problems. Wei and Kuo(1969) gave an exact lumping analysis of a unimolecular reaction system coupled with diffusion under steady-state conditions. In the present paper a general lumping analysis of an arbitrary reaction system coupled with diffusion under both steady-state and transient conditions is presented. It will be shown that similar results to those of the non-diffusion reaction systems can be obtained. Section II discusses exact lumping for a steady-state condition. Section III considers exact lumping for the transient condition. Section IV presents the conditions for exact lumping when the diffusivity is a function of position or the concentrations of the reactants. In section V, a discussion of approximate lumping is presented. Finally, Section VI presents the conclusion and discussion of the paper.

II. EXACT LUMPING FOR A REACTION SYSTEM COUPLED WITH DIFFUSION UNDER STEADY-STATE CONDITIONS

Consider an arbitrary complex reaction system with n -species occurring within a porous catalyst particle(Wei, 1962). Other diffusion problems can be treated in the same way. Let V be the interior of the catalyst particle, and ∂V be the boundary of V across which mass transfer may occur. At a point represented by the vector

\mathbf{r} within the catalyst particle, the local reaction rate vector is determined, in terms of the n -dimensional local concentration vector $\mathbf{y}(\mathbf{r})$, by $\mathbf{f}(\mathbf{y}(\mathbf{r}))$ which does not contain \mathbf{r} explicitly. The diffusion rate vector of supply of the species to the point \mathbf{r} is given by $D\nabla^2\mathbf{y}(\mathbf{r})$, where D is the n -dimensional diagonal effective diffusivity matrix with positive number d_i as its i th diagonal element. Here we consider d_i to be independent of concentrations and position. We will discuss the cases when d_i is a function of position or the concentrations of the reactants in Section IV. In a steady-state, at point \mathbf{r} the reaction rate vector must equal the negative rate vector of supply by diffusion

$$-D\nabla^2\mathbf{y}(\mathbf{r}) = \mathbf{f}(\mathbf{y}(\mathbf{r})). \quad \mathbf{r} \in V. \quad (1)$$

We now give the definition of exact lumping validated in the n -dimensional space of $\mathbf{y}(\mathbf{r})$ for a reaction system coupled with diffusion under steady-state conditions. The reaction-diffusion system in Equation 1 is exactly lumpable by an $\hat{n} \times n$ ($\hat{n} \leq n$) constant matrix M with rank \hat{n} if for

$$\hat{\mathbf{y}}(\mathbf{r}) = M\mathbf{y}(\mathbf{r}), \quad (2)$$

we can find an $\hat{n} \times \hat{n}$ nonsingular constant matrix \hat{D} and an \hat{n} -function vector $\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}))$ such that the behavior of $\hat{\mathbf{y}}(\mathbf{r})$ can be described by

$$-\hat{D}\nabla^2\hat{\mathbf{y}}(\mathbf{r}) = \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r})). \quad (3)$$

According to the physical meaning of an effective diffusivity matrix, \hat{D} is a nonsingular constant diagonal matrix with positive diagonal elements. However, sometimes these conditions cannot be satisfied. Our main task is reducing the dimension. Therefore, it is not necessary to satisfy all these restrictions. Here we only constrain \hat{D} to be nonsingular. If \hat{D} is not diagonal with positive diagonal elements, Equation 3 is satisfactory mathematically if not physically.

A. Necessary and Sufficient Conditions for Exact Lumping under Steady-state Conditions

Not every system is exactly lumpable. Therefore, we need to determine the necessary and sufficient conditions for the existence of exact lumping. We also desire that these conditions be constructive in order to determine the lumping matrices. First rewrite Equations 1 and 3 as

$$\nabla^2 \mathbf{y}(\mathbf{r}) = -D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r})), \quad (4)$$

$$\nabla^2 \hat{\mathbf{y}}(\mathbf{r}) = -\hat{D}^{-1} \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r})), \quad (5)$$

and considering Equation 2 we have

$$M \nabla^2 \mathbf{y}(\mathbf{r}) = -\hat{D}^{-1} \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r})). \quad (6)$$

Multiplying both sides of Equation 4 by M from the left gives

$$M \nabla^2 \mathbf{y}(\mathbf{r}) = -M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r})), \quad (7)$$

and upon comparing Equations 6 and 7 we have

$$M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r})) = \hat{D}^{-1} \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r})), \quad (8)$$

$$M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r})) = \hat{D}^{-1} \hat{\mathbf{f}}(M \mathbf{y}(\mathbf{r})). \quad (9)$$

As the rank of M is \hat{n} , there must exist generalized inverses (Ben-Israel and Greville, 1974) \tilde{M} of matrix M satisfying

$$M \tilde{M} = I_{\hat{n}}, \quad (10)$$

where $I_{\hat{n}}$ is the \hat{n} -identity matrix. We consider the lumping problem generally, i.e., the lumping scheme is validated in the whole n -dimensional space of $\mathbf{y}(\mathbf{r})$. Then

Equation 9 is an identity for any $y(r)$. Therefore letting $y(r)$ take the value $\bar{M}\hat{y}(r)$ and substituting it into Equation 9, we have

$$MD^{-1}f(\bar{M}\hat{y}(r)) = \hat{D}^{-1}\hat{f}(M\bar{M}\hat{y}(r)), \quad (11)$$

$$MD^{-1}f(\bar{M}My(r)) = \hat{D}^{-1}\hat{f}(\hat{y}(r)). \quad (12)$$

Comparing Equations 8 and 12, we obtain the necessary condition for the existence of exact lumping

$$MD^{-1}f(y(r)) = MD^{-1}f(\bar{M}My(r)). \quad (13)$$

Equation 13 is also sufficient for the existence of exact lumping. Indeed, if we multiply both sides of Equation 4 from the left by M and utilizing Equation 13, we obtain

$$\begin{aligned} M\nabla^2 y(r) &= \nabla^2 My(r) \\ &= -MD^{-1}f(y(r)) \\ &= -MD^{-1}f(\bar{M}My(r)). \end{aligned} \quad (14)$$

Let

$$\hat{y}(r) = My(r), \quad (15)$$

$$\hat{D}^{-1}\hat{f}(\hat{y}(r)) = MD^{-1}f(\bar{M}\hat{y}(r)). \quad (16)$$

Then Equation 14 becomes

$$\nabla^2 \hat{y}(r) = -\hat{D}^{-1}\hat{f}(\hat{y}(r)). \quad (17)$$

Multiplying both sides of the above equation from the left by $-\hat{D}$ yields Equation 3. This shows that the system of Equation 1 is exactly lumpable by M . Considering Equation 16, the lumped system can then be described as follows:

$$-\hat{D}\nabla^2 \hat{y}(r) = \hat{D}MD^{-1}f(\bar{M}\hat{y}(r)). \quad (18)$$

Notice that \hat{D} can be chosen arbitrarily, except that it is nonsingular. Considering the physical meaning of effective diffusivity matrix, we would like \hat{D} to be a nonsingular constant diagonal matrix with positive diagonal elements. The simplest case is that $\hat{D} = I_n$. In this case Equation 18 becomes

$$-\nabla^2 \hat{y}(\mathbf{r}) = MD^{-1}f(\bar{M}\hat{y}(\mathbf{r})). \quad (19)$$

Equation 13 does not place any restriction on \bar{M} except that $M\bar{M} = I_n$. The latter point is important in that the non-unique nature of \bar{M} does not effect the form of the lumped equations in the exact case. This means that \bar{M} in Equation 18 is anyone of the generalized inverses satisfying $M\bar{M} = I_n$. This can be easily demonstrated as follows.

Considering once again that Equation 13 is an identity for all $y(\mathbf{r})$, let $y(\mathbf{r})$ take the following value

$$\bar{M}'My(\mathbf{r}),$$

where \bar{M}' is another generalized inverse of M . We obtain

$$\begin{aligned} MD^{-1}f(\bar{M}'My(\mathbf{r})) &= MD^{-1}f(\bar{M}M\bar{M}'My(\mathbf{r})), \\ &= MD^{-1}f(\bar{M}My(\mathbf{r})), \end{aligned} \quad (20)$$

or

$$MD^{-1}f(\bar{M}'\hat{y}(\mathbf{r})) = MD^{-1}f(\bar{M}\hat{y}(\mathbf{r})). \quad (21)$$

This shows that different generalized inverses of M give the same lumped model.

We cannot directly apply Equation 13 to examine whether a system is exactly lumpable or not, because we do not know M in advance. In order to obtain further insight into exact lumping, we differentiate both sides of Equation 13 with respect to $y(\mathbf{r})$ to produce

$$MD^{-1}J(y(\mathbf{r})) = MD^{-1}J(\bar{M}My(\mathbf{r}))\bar{M}M. \quad (22)$$

Equation 22 is not only the necessary condition for exact lumping, but the sufficient one as well. Integrating Equation 22 under an appropriate integration condition with respect to $\mathbf{y}(\mathbf{r})$ will yield Equation 13, which is the necessary and sufficient condition for the existence of exact lumping. Since the rank of M is \hat{n} , it has a nontrivial null space \mathcal{N} with dimension $n - \hat{n}$. We can verify that \mathcal{N} is invariant under $D^{-1}J(\mathbf{y}(\mathbf{r}))$, no matter what value $\mathbf{y}(\mathbf{r})$ takes. Indeed, for every $\mathbf{x} \in \mathcal{N}$ we have

$$MD^{-1}J(\mathbf{y}(\mathbf{r}))\mathbf{x} = MD^{-1}J(\tilde{M}M\mathbf{y}(\mathbf{r}))\tilde{M}M\mathbf{x} = 0. \quad (23)$$

This implies that $D^{-1}J(\mathbf{y}(\mathbf{r}))\mathbf{x} \in \mathcal{N}$ for any value of $\mathbf{y}(\mathbf{r})$, so \mathcal{N} is $D^{-1}J(\mathbf{y}(\mathbf{r}))$ -invariant.

Suppose \mathcal{N} is represented as

$$\mathcal{N} = \text{Span}\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-\hat{n}}\}, \quad (24)$$

where \mathbf{x}_i 's are the basis of \mathcal{N} . Let vectors \mathbf{x}_i compose the columns of matrix X , then

$$MX = 0, \quad (25)$$

and

$$MD^{-1}J(\mathbf{y}(\mathbf{r}))X = 0. \quad (26)$$

Notice that if \mathcal{N} is $D^{-1}J(\mathbf{y}(\mathbf{r}))$ -invariant, then \mathcal{N}^\perp is $J^T(\mathbf{y}(\mathbf{r}))(D^{-1})^T$ -invariant. Since D^{-1} is diagonal, \mathcal{N}^\perp is also $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant (Gohberg et al., 1986). Let $\mathcal{M} = \mathcal{N}^\perp$. Considering Equation 25, it is obvious that \mathcal{M} is spanned by the row vectors of M .

$$\mathcal{M} = \text{Span}\{\mathbf{m}_{(1)}, \mathbf{m}_{(2)}, \dots, \mathbf{m}_{(\hat{n})}\}, \quad (27)$$

where $\mathbf{m}_{(i)}$ is the transpose of row i of M . We call \mathcal{N} and \mathcal{M} fixed invariant subspaces of $J(\mathbf{y}(\mathbf{r}))$ and $J^T(\mathbf{y}(\mathbf{r}))$, respectively.

In conclusion, a system described as Equation 1 can be exactly lumped by an $\hat{n} \times n$ real constant matrix M , only if the subspace \mathcal{M} spanned by the row vectors of M is $J^T(y(r))D^{-1}$ -invariant. We can demonstrate that this condition is also sufficient and the lumped model can be represented as

$$\nabla^2 \hat{y}(r) = -MD^{-1}f(\bar{M}\hat{y}(r)). \quad (28)$$

The proofs are given in Appendix.

Similarly, as the non-diffusion system we also have the following equation for an exactly lumpable reaction-diffusion system(Li and Rabitz, 1989):

$$M(D^{-1}J(y(r)) - D^{-1}J(\bar{M}My(r))) = 0. \quad (29)$$

This equation implies that $J^T(y(r))D^{-1}$ and $J^T(\bar{M}My(r))D^{-1}$ have the same eigenvalues corresponding to \mathcal{M} .

As a special case, when a system is linear, i.e., unimolecular, $J(y(r))$ is a constant matrix and so is $J^T(y(r))D^{-1}$. In this situation, the fixed invariant subspaces become the invariant subspaces of a constant matrix and do exist. Therefore, a linear system is always exactly lumpable.

Similarly, as the non-diffusion system, when a fixed $J^T(y(r))D^{-1}$ -invariant subspace corresponds to constant eigenvalues, the lumped system is linear, no matter if the original system is linear or not(Li and Rabitz, 1989).

In summary, for exact lumping in the whole n -dimensional composition space for a reaction-diffusion system under steady-state conditions we need to determine whether the fixed nontrivial invariant subspaces \mathcal{M} of $J^T(y(r))D^{-1}$ exist or not. If they do exist, the system described as Equation 1 is exactly lumpable by matrix M , whose rows are composed of the basis vectors of \mathcal{M} . The lumped system can be described by Equation 18.

B. Determination of the Fixed $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant Subspaces \mathcal{M}

In order to determine lumping matrices M we need first to determine the fixed $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant subspaces \mathcal{M} . As we have proved in our previous paper (Li and Rabitz, 1989), $J^T(\mathbf{y}(\mathbf{r}))$ can be represented as a linear combination of $m(m \leq n^2)$ constant matrices

$$J^T(\mathbf{y}(\mathbf{r})) = \sum_{k=1}^m a_k(\mathbf{y}(\mathbf{r}))A_k, \quad (30)$$

where $a_k(\mathbf{y}(\mathbf{r}))$ are parameters which are functions of $\mathbf{y}(\mathbf{r})$; the A_k 's are constant matrices considered as a basis of $J^T(\mathbf{y}(\mathbf{r}))$. Then we have

$$\begin{aligned} J^T(\mathbf{y}(\mathbf{r}))D^{-1} &= \sum_{k=1}^m a_k(\mathbf{y}(\mathbf{r}))A_k D^{-1} \\ &= \sum_{k=1}^m a_k(\mathbf{y}(\mathbf{r}))B_k, \end{aligned} \quad (31)$$

where

$$B_k = A_k D^{-1}. \quad (32)$$

It has been demonstrated that the simultaneously invariant subspaces of all the constant matrices A_k are $J^T(\mathbf{y}(\mathbf{r}))$ -invariant (Li and Rabitz, 1989). Similarly, the simultaneously invariant subspaces of all the constant matrices B_k are $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant.

When a reaction system is uni- and/or bimolecular, the elements of $J^T(\mathbf{y}(\mathbf{r}))$ are only linear functions of the $y_k(\mathbf{r})$'s. In this case, Equation 30 will have a simple form, i.e., $a_k(\mathbf{y}(\mathbf{r}))$ is either constant or $y_k(\mathbf{r})$:

$$J^T(\mathbf{y}(\mathbf{r})) = A_0 + \sum_{k=1}^m y_k(\mathbf{r})A_k, \quad (33)$$

where m is equal to or less than n , and A_0 can be the null matrix. In this case the fixed $J^T(y(r))$ -invariant subspaces are simultaneously A_0 - and all A_k -invariant. Similarly, we also have

$$J^T(y(r))D^{-1} = B_0 + \sum_{k=1}^m y_k(r)B_k, \quad (34)$$

and the fixed $J^T(y(r))D^{-1}$ -invariant subspaces are simultaneously B_0 - and all B_k -invariant. Therefore, we can determine the fixed invariant subspaces of $J^T(y(r))D^{-1}$ by determining the simultaneously invariant ones of all $B_k (k = 0, 1, \dots, m)$. The procedure to determine the simultaneously invariant subspaces of all B_k through $\text{Inv}(\sum_{k=0}^m B_k)$ or $\text{Inv}(\prod_{k=0}^m B_k)$ has been given in a previous paper (Li and Rabitz, 1989).

Let us consider a special case that there exist simultaneously all A_k and D^{-1} invariant subspaces \mathcal{M} . We can prove that \mathcal{M} are simultaneously all B_k -invariant. Indeed, for any $x \in \mathcal{M}$, we have

$$B_k x = A_k D^{-1} x = A_k x' = x'' \in \mathcal{M}, \quad (35)$$

where $x' \in \mathcal{M}$ because \mathcal{M} is D^{-1} -invariant.

We can prove that any D^{-1} -invariant subspace is also D -invariant. Since D^{-1} is nonsingular, any invariant subspace \mathcal{M} of it is a nonsingular invariant one, i.e., the image of \mathcal{M} upon mapping by D^{-1} has the same dimension as that of \mathcal{M} . In this case, its corresponding matrix representation M^T satisfies the following equation

$$D^{-1} M^T = M^T Q^{-1}, \quad (36)$$

where Q^{-1} is an $\hat{n} \times \hat{n}$ nonsingular matrix. Multiplying both sides of Equation 36 from the left and right by D and Q , respectively, yields

$$DD^{-1} M^T Q = DM^T Q^{-1} Q, \quad (37)$$

$$M^T Q = D M^T. \quad (38)$$

This implies that \mathcal{M} is D -invariant. Transposing Equation 38 gives

$$Q^T M = M D^T = M D. \quad (39)$$

Under this condition the exact lumping problem of a reaction system coupled with diffusion becomes simple. Suppose \mathcal{M} is simultaneously all A_k and D -invariant, i.e., simultaneously $J^T(\mathbf{y}(\mathbf{r}))$ - and D -invariant. In this case, from the result obtained in our previous paper of exact lumping for the non-diffusion system, we have

$$M f(\bar{M} M \mathbf{y}(\mathbf{r})) = M f(\mathbf{y}(\mathbf{r})). \quad (40)$$

Multiplying both sides of Equation 1 from the left by M gives

$$\begin{aligned} -M D \nabla^2 \mathbf{y}(\mathbf{r}) &= M f(\mathbf{y}(\mathbf{r})), \\ -Q^T M \nabla^2 \mathbf{y}(\mathbf{r}) &= M f(\bar{M} M \mathbf{y}(\mathbf{r})). \end{aligned} \quad (41)$$

Let

$$\hat{\mathbf{y}}(\mathbf{r}) = M \mathbf{y}(\mathbf{r}).$$

Then we have

$$-Q^T \nabla^2 \hat{\mathbf{y}}(\mathbf{r}) = M f(\bar{M} \hat{\mathbf{y}}(\mathbf{r})). \quad (42)$$

We can see that the system is exactly lumpable by M , and Q^T is just like \hat{D} . Considering Equation 39 we have

$$Q^T = M D \bar{M}, \quad (43)$$

which may not be diagonal. Since Q is nonsingular, if we require \hat{D} to be a nonsingular diagonal matrix, we can multiply both sides of Equation 42 from the left by $(Q^T)^{-1}$ to produce

$$\begin{aligned}
-\nabla^2 \hat{y}(\mathbf{r}) &= (Q^T)^{-1} M f(\bar{M} \hat{y}(\mathbf{r})) \\
&= (Q^{-1})^T M f(\bar{M} \hat{y}(\mathbf{r})) \\
&= M D^{-1} f(\bar{M} \hat{y}(\mathbf{r})).
\end{aligned} \tag{44}$$

Here we have used the relation of Equation 36. Then we can multiply both sides of Equation 44 from the left by an arbitrary nonsingular diagonal matrix \hat{D} to obtain the standard form

$$-\hat{D} \nabla^2 \hat{y}(\mathbf{r}) = \hat{D} M D^{-1} f(\bar{M} \hat{y}(\mathbf{r})). \tag{45}$$

When $L = dI_n$, where d is a positive number, the lumping problem is even simpler. Since any subspace is dI_n -invariant, therefore the necessary and sufficient condition is reduced to the condition of the non-diffusion system:

$$MJ(y(\mathbf{r})) = MJ(\bar{M} M y(\mathbf{r})) \bar{M} M. \tag{46}$$

In this case, Equation 18 becomes

$$-\hat{D} \nabla^2 \hat{y}(\mathbf{r}) = \frac{1}{d} \hat{D} M f(\bar{M} \hat{y}(\mathbf{r})). \tag{47}$$

C. Sample Problem

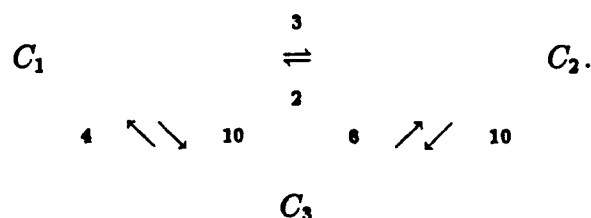
As an example of the application of the analysis above, we choose the simplest case of a unimolecular reaction system. For a unimolecular reaction system, the corresponding differential equations are

$$-D \nabla^2 y(\mathbf{r}) = K y(\mathbf{r}), \tag{48}$$

where K is the rate constant matrix. The Jacobian matrix for $f(y(\mathbf{r}))$ is just K , and then

$$J^T(y(\mathbf{r})) = K^T. \tag{49}$$

As a specific illustration consider a unimolecular reaction system with 3 species (Wei and Kuo, 1969) coupled with diffusion:



where C_1, C_2 and C_3 represent the three species; all numbers are unitless rate constants. Let y_i represent the concentration of species C_i . Then

$$J^T(y(r)) = K^T = \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix}^T. \quad (50)$$

The effective diffusivity matrix is given as

$$D = \begin{pmatrix} 2 & & \\ & 2 & \\ & & 1 \end{pmatrix}. \quad (51)$$

From Section IIA we know that any linear system coupled with diffusion under steady-state conditions is exactly lumpable. Then the only thing we need to do is to determine all of the $K^T D^{-1}$ -invariant subspaces, whose basis vectors compose the lumping matrices.

$$\begin{aligned}
 K^T D^{-1} &= \begin{pmatrix} -13 & 3 & 10 \\ 2 & -12 & 10 \\ 4 & 6 & -10 \end{pmatrix} \begin{pmatrix} 0.5 & & \\ & 0.5 & \\ & & 1 \end{pmatrix} \\
 &= \begin{pmatrix} -6.5 & 1.5 & 10 \\ 1 & -6 & 10 \\ 2 & 3 & -10 \end{pmatrix}. \quad (52)
 \end{aligned}$$

The eigenvector matrix X and the eigenvalue matrix Λ of $K^T D^{-1}$ are

$$X = \begin{pmatrix} 1 & 1 & 1 \\ -2/3 & 1 & 1 \\ 0 & -1 & 1/2 \end{pmatrix}, \quad (53)$$

$$\Lambda = \begin{pmatrix} -15/2 & & \\ & -15 & \\ & & 0 \end{pmatrix}. \quad (54)$$

Considering that the eigenvalues of $K^T D^{-1}$ are distinct, any subspace spanned by a subset of its eigenvectors is invariant to it. For convenience let x_1, x_2 and x_3 represent the 3 columns of X . Then the set of all $K^T D^{-1}$ -invariant subspaces $\text{Inv}(K^T D^{-1})$ contains

$$\begin{aligned} &\text{Span}\{0\}, \text{Span}\{x_1\}, \text{Span}\{x_2\}, \text{Span}\{x_3\}, \\ &\text{Span}\{x_1, x_2\}, \text{Span}\{x_1, x_3\}, \text{Span}\{x_2, x_3\}, \\ &\quad \mathcal{R}^3 \end{aligned}$$

In $\text{Inv}(K^T D^{-1})$ the nontrivial invariant subspaces, i.e., those with dimension 1 and 2, can be used to construct the exact lumping matrices. Choosing some bases for the nontrivial invariant subspaces \mathcal{M} the corresponding lumping matrices are as follows:

The lumping matrices for 1-dimensional \mathcal{M} :

$$M_1 = \begin{pmatrix} 1 & -2/3 & 0 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} 1 & 1 & -1 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} 1 & 1 & 1/2 \end{pmatrix}.$$

The lumping matrices for 2-dimensional \mathcal{M} :

$$M_4 = \begin{pmatrix} 1 & -2/3 & 0 \\ 1 & 1 & -1 \end{pmatrix},$$

$$M_5 = \begin{pmatrix} 1 & -2/3 & 0 \\ 1 & 1 & 1/2 \end{pmatrix},$$

$$M_6 = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & 1/2 \end{pmatrix}.$$

The number of $K^T D^{-1}$ -invariant subspaces is finite, but the number of the lumping matrices is infinite, because one can choose different bases to represent 2-dimensional invariant subspaces. For example, $\text{Span}\{x_2, x_3\}$ gives the lumping

matrix M_6 . We can use elementary row operations (Lang, 1986) on the two rows to produce another equivalent exact lumping matrix:

$$M_7 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The rows of the new lumping matrix are just another basis of the same invariant subspace.

In Section IIB we proved that the simultaneously D - and $J^T(\mathbf{y}(\mathbf{r}))$ -invariant subspaces are contained in $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant ones. Here this means that the simultaneously D - and K^T -invariant subspaces are contained in the $K^T D^{-1}$ -invariant ones. To show this we determine the simultaneously D - and K^T -invariant subspaces, which are contained in the invariant ones of matrix $A = D + K^T$.

$$A = \begin{pmatrix} -11 & 3 & 10 \\ 2 & -10 & 10 \\ 4 & 6 & -9 \end{pmatrix}. \quad (55)$$

The eigenvector matrix X and the eigenvalue matrix Λ of A are

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -2/3 \\ -(1 + \sqrt{401})/20 & (-1 + \sqrt{401})/20 & 0 \end{pmatrix}, \quad (56)$$

$$\Lambda = \begin{pmatrix} -(17 + \sqrt{401})/20 & & \\ & (-17 + \sqrt{401})/20 & \\ & & -13 \end{pmatrix}. \quad (57)$$

Since all eigenvalues of A are distinct, similarly all the subspaces spanned by the subset of the eigenvectors are A -invariant. After examining which of A -invariant subspaces are simultaneously D - and K^T -invariant, we obtain the simultaneously D - and K^T -invariant subspaces, whose matrix representations are as follows:

The matrix representation for 1-dimensional \mathcal{M} :

$$M_8 = \begin{pmatrix} 1 & -2/3 & 0 \end{pmatrix}.$$

The matrix representation for 2-dimensional \mathcal{M} :

$$M_9 = \begin{pmatrix} 1 & 1 & -(1 + \sqrt{401})/20 \\ 1 & 1 & (-1 + \sqrt{401})/20 \end{pmatrix}.$$

One can see that $M_8 = M_1$ and M_9 , just like M_7 , is only another matrix representation of the corresponding subspace for M_8 . This result shows that the simultaneously D - and K^T -invariant subspaces are really contained in $K^T D^{-1}$ -invariant ones.

In Section IIA we proved that the non-unique nature of \bar{M} does not effect the form of the lumped equations. To illustrate this point consider for M_1 , for example, where we can find an infinite number of \bar{M}_1 satisfying $M_1 \bar{M}_1 = 1$. We arbitrarily choose three:

$$\bar{M}_1 = (1 \ 0 \ 0)^T, \quad \bar{M}_1 = (1 \ 0 \ 1)^T, \quad \bar{M}_1 = (5/3 \ 1 \ 0)^T.$$

It is easy to show that the differential equations for the lumped system are independent on the choice of \bar{M}_1 . According to Equation 18 and letting $\hat{D} = I_A$ we have

$$-\nabla^2 \hat{y}(\mathbf{r}) = M D^{-1} f(\bar{M} \hat{y}(\mathbf{r})), \quad (58)$$

and since

$$f(\mathbf{y}(\mathbf{r})) = K \mathbf{y}(\mathbf{r}), \quad (59)$$

then we have

$$-\nabla^2 \hat{y}(\mathbf{r}) = M D^{-1} K \bar{M} \hat{y}(\mathbf{r}). \quad (60)$$

It is easy to verify that for different \bar{M}_1 we have the same lumped equation:

$$\begin{aligned} -\nabla^2 \hat{y}(\mathbf{r}) &= (1 \ -2/3 \ 0) \begin{pmatrix} 1/2 & & \\ & 1/2 & \\ & & 1 \end{pmatrix} \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \bar{M}_1 \hat{y}(\mathbf{r}) \\ &= (-15/2 \ 5 \ 0) \bar{M}_1 \hat{y}(\mathbf{r}) \\ &= -\frac{15}{2} M_1 \bar{M}_1 \hat{y}(\mathbf{r}) \\ &= -\frac{15}{2} \hat{y}(\mathbf{r}). \end{aligned} \quad (61)$$

Similarly we can obtain the lumped equations for other lumping matrices M_2 to M_7 as follows:

$$\nabla^2 \hat{y}(\mathbf{r}) = 15 \hat{y}(\mathbf{r}). \quad (62)$$

$$\nabla^2 \hat{y}(\mathbf{r}) = 0. \quad (63)$$

$$\nabla^2 \hat{y}(\mathbf{r}) = \begin{pmatrix} 15/2 & \\ & 15 \end{pmatrix} \hat{y}(\mathbf{r}). \quad (64)$$

$$\nabla^2 \hat{y}(\mathbf{r}) = \begin{pmatrix} 15/2 & \\ & 0 \end{pmatrix} \hat{y}(\mathbf{r}). \quad (65)$$

$$\nabla^2 \hat{y}(\mathbf{r}) = \begin{pmatrix} 15 & \\ & 0 \end{pmatrix} \hat{y}(\mathbf{r}). \quad (66)$$

$$\nabla^2 \hat{y}(\mathbf{r}) = \begin{pmatrix} 5 & \\ & 10 \end{pmatrix} \hat{y}(\mathbf{r}). \quad (67)$$

III. EXACT LUMPING FOR A REACTION SYSTEM COUPLED WITH DIFFUSION UNDER TRANSIENT CONDITIONS

A. Necessary and Sufficient Conditions for Exact Lumping under Transient Conditions

As a reasonable assumption we take that the ambient concentration vector $\mathbf{y}(\mathbf{R}, t)$ ($\mathbf{R} \in \partial V$) does not change with time and that the concentration vector $\mathbf{y}(\mathbf{r}, t)$ in the interior of the catalyst particle is initially zero. The differential equations corresponding to transient conditions are as follows:

$$\frac{\partial}{\partial t} \mathbf{y}(\mathbf{r}, t) - D \nabla^2 \mathbf{y}(\mathbf{r}, t) - \mathbf{f}(\mathbf{y}(\mathbf{r}, t)) = 0, \quad (68)$$

where the first term on the left side represents the accumulation of the reactants due to diffusion and reactions.

The definition of exact lumping validated in the n -dimensional composition space under transient conditions can be given as follows. If a reaction system coupled with diffusion under transient conditions described as Equation 68 can be exactly lumped by an $\hat{n} \times n$ constant matrix M with rank \hat{n} , this means that for

$$\hat{\mathbf{y}}(\mathbf{r}, t) = M\mathbf{y}(\mathbf{r}, t), \quad (69)$$

we can find an $\hat{n} \times \hat{n}$ nonsingular constant matrix \hat{D} and an \hat{n} -function vector $\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t))$ such that the behavior of $\hat{\mathbf{y}}(\mathbf{r}, t)$ can be described by

$$\frac{\partial}{\partial t} \hat{\mathbf{y}}(\mathbf{r}, t) - \hat{D} \nabla^2 \hat{\mathbf{y}}(\mathbf{r}, t) - \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t)) = 0. \quad (70)$$

As discussed in the previous section, here we only constrain \hat{D} to be nonsingular.

Equation 70 is valid for any value of t including $t \rightarrow \infty$, i.e., a steady-state. In a steady-state, the first term vanishes and Equation 70 becomes Equation 3. From Equations 14 and 16 we know that

$$\begin{aligned} \lim_{t \rightarrow \infty} \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t)) &= \lim_{t \rightarrow \infty} \hat{D} M D^{-1} \mathbf{f}(M \hat{\mathbf{y}}(\mathbf{r}, t)) \\ &= \lim_{t \rightarrow \infty} \hat{D} M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r}, t)). \end{aligned} \quad (71)$$

Notice that $\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t))$ and $\mathbf{f}(\mathbf{y}(\mathbf{r}, t))$ are only explicit functions of $\hat{\mathbf{y}}$ and \mathbf{y} , and do not contain \mathbf{r}, t explicitly. Therefore, $\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t))$ must have the same form in Equations 70 and 71. Otherwise, the lumped scheme in the transient regime cannot be validated in the steady-state condition. The only difference is that $\hat{\mathbf{y}}$ is a function of \mathbf{r} and t in Equation 70 instead of a function of only \mathbf{r} in Equation 70. Then we have

$$\begin{aligned} \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t)) &= \hat{D} M D^{-1} \mathbf{f}(\bar{M} \hat{\mathbf{y}}(\mathbf{r}, t)) \\ &= \hat{D} M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r}, t)). \end{aligned} \quad (72)$$

Considering this point and Equation 69, Equation 70 can be rewritten as

$$M \frac{\partial}{\partial t} \mathbf{y}(\mathbf{r}, t) - \hat{D} M \nabla^2 \mathbf{y}(\mathbf{r}, t) - \hat{D} M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r}, t)) = 0. \quad (73)$$

Now we need to determine the condition under which a system coupled with diffusion under transient conditions is exactly lumpable. Multiplying Equation 68 from the left by $\hat{D} M D^{-1}$ yields

$$\hat{D} M D^{-1} \frac{\partial}{\partial t} \mathbf{y}(\mathbf{r}, t) - \hat{D} M \nabla^2 \mathbf{y}(\mathbf{r}, t) - \hat{D} M D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r}, t)) = 0. \quad (74)$$

Subtracting Equation 73 from Equation 74 gives

$$(\hat{D} M D^{-1} - M) \frac{\partial}{\partial t} \mathbf{y}(\mathbf{r}, t) = 0. \quad (75)$$

This equation holds for any value of $\partial \mathbf{y}(\mathbf{r}, t) / \partial t$. Considering Equation 68 we have

$$\frac{\partial}{\partial t} \mathbf{y}(\mathbf{r}, t) = D \nabla^2 \mathbf{y}(\mathbf{r}, t) + \mathbf{f}(\mathbf{y}(\mathbf{r}, t)). \quad (76)$$

Notice that D is a nonsingular matrix and in realistic problems the diffusivities for different species are usually different. Therefore, we can choose different initial values of $\mathbf{y}(\mathbf{R}, 0)$ so that $\partial \mathbf{y}(\mathbf{r}, t) / \partial t$ can be an arbitrary vector in n -dimensional space. Under this condition, Equation 75 is valid only if

$$\hat{D} M D^{-1} - M = 0. \quad (77)$$

This relation is equivalent to

$$\hat{D} M = M D, \quad (78)$$

or considering that $D^T = D$ we have

$$M^T \hat{D}^T = D M^T. \quad (79)$$

This equation shows that the subspace \mathcal{M} spanned by the row vectors of M is D -invariant.

According to the result obtained above the necessary condition for exact lumping of a reaction-diffusion system under transient conditions is Equations 72 and 77 (or 78, 79). Notice that utilizing Equation 77 we can represent Equation 72 as

$$\begin{aligned}\hat{f}(\hat{y}(r,t)) &= Mf(\bar{M}\hat{y}(r,t)) \\ &= Mf(y(r,t)).\end{aligned}\tag{80}$$

It is easy to demonstrate that this condition is also sufficient for the existence of exact lumping of a reaction system coupled with diffusion under transient conditions. Multiplying Equation 68 from the left by M yields

$$M \frac{\partial}{\partial t} y(r,t) - MD \nabla^2 y(r,t) - Mf(y(r,t)) = 0.$$

Letting $\hat{y}(r,t) = My(r,t)$ and substituting Equations 78 and 80 into the above equation gives

$$\frac{\partial}{\partial t} \hat{y}(r,t) - \hat{D} \nabla^2 \hat{y}(r,t) - \hat{f}(\hat{y}(r,t)) = 0.\tag{81}$$

Then letting

$$\hat{f}(\hat{y}(r,t)) = Mf(\bar{M}\hat{y}(r,t)),$$

we have

$$\frac{\partial}{\partial t} \hat{y}(r,t) - \hat{D} \nabla^2 \hat{y}(r,t) - \hat{f}(\hat{y}(r,t)) = 0.$$

This is Equation 70.

From the results of Section IIB, Equations 79 and 80 imply that $J^T(y(r,t))$ and D have simultaneously invariant subspaces \mathcal{M} . Thus we obtain the conclusion: A reaction-diffusion system under transient conditions is exactly lumpable if and only if there exist simultaneously nontrivial fixed $J^T(y(r,t))$ - and D -invariant subspaces \mathcal{M} . The lumping matrices M are the matrix representations of \mathcal{M} .

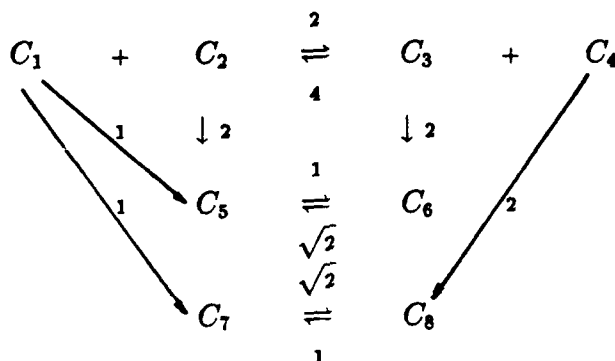
Notice that in this case we can no longer choose \hat{D} arbitrarily. According to Equation 78 we have

$$\hat{D} = MDM^T. \quad (82)$$

The resultant \hat{D} may not be diagonal.

B. Sample Problem

As an example of the application of the analysis above, we choose the uni- and bimolecular reaction system used in our previous paper. A uni- and bimolecular reaction system with 8 species (Li, 1984) is illustrated as follows:



where the C_i 's are species; the numbers are unitless rate constants.

Letting y_i represent the concentration of C_i , it is easy to write out the kinetic equations and the transpose of the corresponding Jacobian matrix $J^T(y(r,t))$.

$$\begin{aligned}
 dy_1/dt &= -2y_1 - 2y_1y_2 + 4y_3y_4 \\
 dy_2/dt &= -2y_2 - 2y_1y_2 + 4y_3y_4 \\
 dy_3/dt &= -2y_3 - 4y_3y_4 + 2y_1y_2 \\
 dy_4/dt &= -2y_4 - 4y_3y_4 + 2y_1y_2 \\
 dy_5/dt &= -y_5 + y_1 + 2y_2 + \sqrt{2}y_6 \\
 dy_6/dt &= -\sqrt{2}y_6 + 2y_3 + y_5 \\
 dy_7/dt &= -\sqrt{2}y_7 + y_1 + y_8 \\
 dy_8/dt &= -y_8 + 2y_4 + \sqrt{2}y_7
 \end{aligned} \quad (83)$$

$$J^T(y(r,t)) = \begin{pmatrix} -2(1+y_2) & -2y_2 & 2y_2 & 2y_2 & 1 & 0 & 1 & 0 \\ -2y_1 & -2(1+y_1) & 2y_1 & 2y_1 & 2 & 0 & 0 & 0 \\ 4y_4 & 4y_4 & -2(1+2y_4) & -4y_4 & 0 & 2 & 0 & 0 \\ 4y_3 & 4y_3 & -4y_3 & -2(1+2y_3) & 0 & 0 & 0 & 2 \\ & 0 & & & -1 & 1 & 0 & 0 \\ & & & & \sqrt{2} & -\sqrt{2} & 0 & 0 \\ & & & & 0 & 0 & -\sqrt{2} & \sqrt{2} \\ & & & & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Suppose that the effective diffusivity matrix D of the system is the following:

$$D = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 2 & & & \\ & & & & & 2 & & \\ & & & & & & 3 & \\ & & & & & & & 3 \end{pmatrix}. \quad (84)$$

We have obtained all the fixed $J^T(y(r,t))$ -invariant subspaces (Li and Rabitz, 1989). The root subspaces of D are

$$\text{Span}\{e_1, e_2, e_3, e_4\}, \text{Span}\{e_5, e_6\}, \text{Span}\{e_7, e_8\}.$$

Any subspace of these root ones and any sum of these subspaces are D -invariant. Then examining which $J^T(y(r,t))$ -invariant subspaces are D -invariant, we obtain the simultaneously D - and fixed $J^T(y(r,t))$ -invariant subspaces. They can be used to construct the exact lumping matrices.

The lumping matrices for 1-dimensional \mathcal{M} :

$$M_1 = (\alpha_1 + \alpha_2 \quad \alpha_3 \quad \alpha_2 \quad \alpha_1 + \alpha_3 \quad 0 \quad 0 \quad 0 \quad 0),$$

The lumping matrices for 2-dimensional \mathcal{M} :

$$M_2 = \begin{pmatrix} \alpha_1 + \alpha_2 & \alpha_3 & \alpha_2 & \alpha_1 + \alpha_3 & 0 & 0 & 0 & 0 \\ \beta_1 + \beta_2 & \beta_3 & \beta_2 & \beta_1 + \beta_3 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $\alpha_i, \beta_i \in \mathcal{R}$. If a matrix contains the same number of α_i 's and β_i 's, the vectors α and β are linearly independent.

The lumping matrices for 3-dimensional \mathcal{M} :

$$M_3 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The lumping matrices for 4-dimensional \mathcal{M} :

$$M_4 = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 0 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix}.$$

The lumping matrices for 5-dimensional \mathcal{M} :

$$M_5 = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix},$$

$$M_6 = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix},$$

$$M_8 = \begin{pmatrix} 1 & & & & & & & \\ & 1 & & & & & & \\ & & 1 & & & & & \\ & & & 1 & & & & \\ & & & & 1 & & & \\ & & & & & 1 & & \\ & & & & & & 1 & \\ & & & & & & & 1 \end{pmatrix}.$$

The lumping matrices for 6-dimensional \mathcal{M} :

$$M_9 = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & -\sqrt{2} \\ & & & & 0 & 0 \\ & & & & 0 & -\sqrt{2} \\ & & & & & 1 \end{pmatrix},$$

$$M_{10} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & 1 \\ & & & & 0 & 0 \\ & & & & 0 & 0 \\ & & & & 1 & 1 \end{pmatrix},$$

$$M_{11} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & 1 \\ & & & & 0 & 0 \\ & & & & 0 & -\sqrt{2} \\ & & & & & 1 \end{pmatrix},$$

$$M_{12} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & -\sqrt{2} \\ & & & & 0 & 0 \\ & & & & 0 & 0 \\ & & & & 1 & 1 \end{pmatrix},$$

$$M_{13} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \\ & & & & 1 & 0 \\ & & & & 0 & 0 \end{pmatrix},$$

$$M_{14} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \\ & & & & 0 & 0 \\ & & & & 0 & 0 \\ & & & & 1 & 0 \\ & & & & 0 & 1 \end{pmatrix}.$$

The lumping matrices for 7-dimensional \mathcal{M} :

$$M_{15} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & 1 \end{pmatrix},$$

$$M_{16} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & \\ & & & & & 1 & \\ & & & & & & -\sqrt{2} \end{pmatrix},$$

$$M_{17} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 0 & 0 & 1 \\ & & & & 0 & 0 & 0 \\ & & & & 1 & 1 & 0 \end{pmatrix},$$

$$M_{18} = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & 1 & & & \\ & & & & 0 & 0 & 1 \\ & & & & 0 & 0 & 0 \\ & & & & 1 & -\sqrt{2} & 0 \end{pmatrix}.$$

The differential equations for the lumped systems can be readily obtained by Equations 80 and 82. For example, for M_1 the lumped equation is

$$\frac{\partial}{\partial t} \hat{y}(\mathbf{r}, t) - \nabla^2 \hat{y}(\mathbf{r}, t) - 2\hat{y}(\mathbf{r}, t) = 0. \quad (85)$$

For M_9 we have

$$\hat{D} = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 2 & \\ & & & & & 3 \end{pmatrix}, \quad (86)$$

and the lumped reaction rate vector $\hat{f}(\hat{y}(r, t))$ is the following:

$$\hat{f}(\hat{y}(r, t)) = \begin{pmatrix} -2\hat{y}_1 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\ -2\hat{y}_2 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\ -2\hat{y}_3 + 2\hat{y}_1\hat{y}_2 - 4\hat{y}_3\hat{y}_4 \\ -2\hat{y}_4 + 2\hat{y}_1\hat{y}_2 - 4\hat{y}_3\hat{y}_4 \\ \hat{y}_1 - 2\hat{y}_2 - 2\sqrt{2}\hat{y}_3 + (1 + \sqrt{2})\hat{y}_5 \\ -\sqrt{2}\hat{y}_1 + 2\hat{y}_4 - (1 + \sqrt{2})\hat{y}_6 \end{pmatrix}. \quad (87)$$

IV. EXACT LUMPING FOR A REACTION SYSTEM WHOSE DIFFUSIVITIES ARE FUNCTIONS OF POSITION OR CONCENTRATIONS

All discussions above are based on the assumption that the diffusivity d_i is independent of position and concentrations. This is true for uniform catalysts and in the Knudsen range. It is also a good approximation for the gaseous diffusion regime. However, when catalysts are not uniform or there are interactions between the diffusion of different species, the diffusivity can be a function of position or concentrations. We will prove that in these cases the sufficient conditions for exact lumping will have similar forms to those already treated.

A. Diffusivity d_i is a Function of Position

Suppose that the diffusivity matrix $D(r)$ is diagonal and a function of r . First we consider the steady-state condition. In this case Equations 1 and 3 become

$$-\nabla D(r) \nabla y(r) = f(y(r)), \quad (88)$$

$$-\nabla \hat{D}(r) \nabla \hat{y}(r) = \hat{f}(\hat{y}(r)), \quad r \in V. \quad (89)$$

In this case it is not easy to determine the necessary condition. However, we can give the sufficient condition of exact lumping in the whole composition space and the desired region of the position vector: $J^T(y(r))$ and $D(r)$ have simultaneously

nontrivial fixed invariant subspaces for all values of $\mathbf{y}(\mathbf{r})$ and \mathbf{r} in the desired region, respectively. The proof is as follows.

When the subspace \mathcal{M} spanned by the row vectors of M is simultaneously $J^T(\mathbf{y}(\mathbf{r}))$ - and $D(\mathbf{r})$ -invariant, as proved before we have

$$Mf(\bar{M}M\mathbf{y}(\mathbf{r})) = Mf(\mathbf{y}(\mathbf{r})). \quad (90)$$

$$MD(\mathbf{r}) = \hat{D}(\mathbf{r})M. \quad (91)$$

Multiplying both sides of Equation 88 from the left by M and using Equations 90 and 91 yields

$$-M\nabla D(\mathbf{r})\nabla\mathbf{y}(\mathbf{r}) = Mf(\mathbf{y}(\mathbf{r})),$$

$$-\nabla MD(\mathbf{r})\nabla\mathbf{y}(\mathbf{r}) = Mf(\bar{M}M\mathbf{y}(\mathbf{r})),$$

$$-\nabla\hat{D}(\mathbf{r})M\nabla\mathbf{y}(\mathbf{r}) = Mf(\bar{M}\hat{\mathbf{y}}(\mathbf{r})),$$

$$-\nabla\hat{D}(\mathbf{r})\nabla\hat{\mathbf{y}}(\mathbf{r}) = \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r})).$$

That is Equation 89.

Under transient conditions and when the diffusivity matrix $D(\mathbf{r})$ is a function of \mathbf{r} , Equations 68 and 70 become

$$\frac{\partial}{\partial t}\mathbf{y}(\mathbf{r},t) - \nabla D(\mathbf{r})\nabla\mathbf{y}(\mathbf{r},t) - \mathbf{f}(\mathbf{y}(\mathbf{r},t)) = \mathbf{0}, \quad (92)$$

$$\frac{\partial}{\partial t}\hat{\mathbf{y}}(\mathbf{r},t) - \nabla\hat{D}(\mathbf{r})\nabla\hat{\mathbf{y}}(\mathbf{r},t) - \hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r},t)) = \mathbf{0}. \quad (93)$$

We can prove that the sufficient condition under steady-state conditions is also sufficient for transient conditions except that $J^T(\mathbf{y}(\mathbf{r}))$ is replaced by $J^T(\mathbf{y}(\mathbf{r},t))$. Since \mathcal{M} is $J^T(\mathbf{y}(\mathbf{r},t))$ -invariant, Equation 90 becomes

$$Mf(\bar{M}M\mathbf{y}(\mathbf{r},t)) = Mf(\mathbf{y}(\mathbf{r},t)). \quad (94)$$

Multiplying Equation 92 from the left by M and using Equations 91 and 94 yields Equation 93.

In conclusion: A reaction-diffusion system with position dependent $D(r)$ under steady-state or transient conditions is exactly lumpable if there exist simultaneously nontrivial fixed $J^T(y(r))$ - and $D(r)$ -invariant subspaces or $J^T(y(r,t))$ - and $D(r)$ -invariant subspaces \mathcal{M} for all values of $y(r)$ and r or $y(r,t)$ and r , respectively. The lumping matrices M are the matrix representations of \mathcal{M} .

B. Diffusivity d_i is a Function of Concentrations of the Reactants

When the diffusivity is dependent on the concentrations of the species in the system, we have not established the necessary condition of exact lumping. The sufficient condition is the same, except that $D(r)$ is replaced by $D(y(r))$ and $D(y(r,t))$ for the steady-state and transient conditions, respectively. In this case Equation 91 becomes

$$MD(y(r)) = \hat{D}(y(r))M \quad (95)$$

and

$$MD(y(r,t)) = \hat{D}(y(r,t))M \quad (96)$$

for the steady-state and transient conditions. The proof is similar.

V. APPROXIMATE LUMPING FOR A REACTION SYSTEM COUPLED WITH DIFFUSION

After we obtain the necessary and sufficient conditions of exact lumping for a reaction system coupled with diffusion under either steady-state or transient conditions, the analysis of approximate lumping for such systems follows by using the results from non-diffusion reaction systems. Here we only discuss the determination of the constrained lumping matrices by the direct approach (Li and Rabitz, 1990b).

The approach of solving the matrix equations to determine the approximate lumping matrices can be treated in the same way. If the given part of the lumping matrix is M_G and $J^T(y)$ has a decomposition as Equation 30, according to the direct approach we need to determine the \hat{n} -dimensional subspace \mathcal{M} containing M_G spanned by the row vectors of M_G . This subspace is as nearly as possible invariant to all the basis constant matrices A_k of $J^T(y)$. The procedure to determine \mathcal{M} is the following. First, we construct a special symmetric matrix Y :

$$Y = \sum_{k=1}^m \sum_{i=0}^{s_k-1} Q(G)_{(ki)}^T Q(G)_{(ki)}, \quad (97)$$

where $Q(G)_{(ki)}^T$ ($k = 1, 2, \dots, m; i = 0, 1, \dots, s_k - 1$) are the orthonormal matrix representations of $\text{Im}(M_G(A_k^T)^i)^T$ and $(M_G(A_k^T)^0)^T = M_G^T$ which can be multiplied by a very large positive number so that M_G^T compose the eigenvectors of Y with the largest eigenvalues. Here s_k is the rank of A_k or is equal to $n - 1$. Second, the eigenvalues and eigenvector matrix R of Y are determined. When the eigenvectors are arranged in R by the nonincreasing order of their eigenvalues, the first \hat{n} eigenvectors form the best constrained approximate lumping matrix containing M_G with row number \hat{n} .

For a reaction-diffusion system under steady-state or transient conditions the exact lumping matrix is related to a subspace \mathcal{M} , which is simultaneously invariant to all constant matrices B_k or all A_k and D , respectively. Therefore, the determination of the constrained approximate lumping matrix is the same as that for a non-diffusion system except that the A_k 's are replaced by B_k 's or A_k 's and D . When D is a function of position or concentrations and $D(r)$, $D(y(r))$ and $D(y(r, t))$ can be decomposed as

$$D(r) = \sum_{i=1}^p b_i(r) D_i, \quad (98)$$

$$D(y(r)) = \sum_{i=1}^q c_i(y(r)) D_i, \quad (99)$$

$$D(y(r,t)) = \sum_{i=1}^r e_i(y(r,t)) D_i, \quad (100)$$

where $b_i(r)$, $c_i(y(r))$ and $e_i(y(r,t))$ are parameters, D_i are constant matrices considered as a basis of $D(r)$, $D(y(r))$ or $D(y(r,t))$. In these cases, the A_k 's are replaced by A_k 's and D_i 's. Then the constrained approximate lumping matrix can be determined in the same way.

VI. CONCLUSION AND DISCUSSION

In this paper a general analysis of exact and approximate lumping for a reaction system coupled with diffusion under both steady-state and transient conditions for constant and position or concentration dependent diffusivity has been given, which can be used for any reaction system. Uni- and/or bimolecular reaction systems are only special cases of this general analysis.

For constant diffusivity, under steady-state conditions the exact lumping matrices can be constructed from the fixed $J^T(y(r))D^{-1}$ -invariant subspaces. The simultaneously D - and $J^T(y(r))$ -invariant subspaces are contained in the set of $J^T(y(r))D^{-1}$ -invariant ones; under transient conditions, the exact lumping matrices are determined by the simultaneously D - and $J^T(y(r,t))$ -invariant subspaces. For position or concentration dependent diffusivity, the sufficient condition is the same as that of the transient regime for constant D except that D is replaced by $D(r)$, $D(y(r))$ or $D(y(r,t))$.

For approximate lumping, the determination of the constrained approximate lumping matrices are almost the same as those of non-diffusion reaction systems. Under steady-state conditions the only difference is that A_k are replaced by $B_k = A_k D^{-1}$. In the transient case the difference is the addition of D . When D is a function of position or concentrations, the basis constant matrices of $D(r)$, $D(y(r))$ or $D(y(r,t))$ are added.

The lumping analysis given above can be further expanded to a more general case. Suppose a system can be described as

$$Ly = f(y), \quad (101)$$

where L is an arbitrary linear operator. The definition of exact lumping of Equation 101 is the following: For

$$\hat{y} = My \quad (102)$$

if we can find an \hat{n} -function vector $\hat{f}(\hat{y})$ such that

$$\hat{L}\hat{y} = \hat{f}(\hat{y}), \quad (103)$$

where \hat{L} is another linear operator satisfying

$$ML = \hat{L}M, \quad (104)$$

we say that Equation 101 is exactly lumpable by M .

According to this definition, one can readily obtain the necessary and sufficient condition of exact lumping for Equation 101 as follows: $J^T(y)$ has nontrivial fixed invariant subspaces. For nondiffusion reaction systems $L = \hat{L} = d/dt$ and Equation 104 always holds. Then the necessary and sufficient condition obtained in our previous work is the same as that of Equation 101. For reaction-diffusion systems Equation 104 gives

$$MD = \hat{D}M. \quad (105)$$

In this case, the necessary and sufficient condition becomes that $J^T(y)$ and D have common fixed invariant subspaces. This is what we obtained for a reaction-diffusion system under transient conditions. It is also sufficient for steady-state conditions.

Since L is an arbitrary linear operator, certain partial differential equation systems belong to Equation 101, and then their lumping problems can be treated.

Equation 101 may be employed to describe an open reaction system in chemical kinetics, mathematical models of some reactors in chemical engineering and a large number of systems in other areas. Therefore, the approaches of exact and approximate lumping developed in our work is quite general and can be used widely.

Acknowledgment

The authors acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

NOTATION

Scalars

- $a_k(y(r))$ = k th coefficient of a linear combination of constant matrices for $J^T(y(r))$
- $b_i(r)$ = parameters of the decomposition of $D(r)$
- $c_i(y(r))$ = parameters of the decomposition of $D(y(r))$
- C_i = i th species of a reaction system
- d = positive constant
- d_i = positive constant
- $e_i(y(r, t))$ = parameters of the decomposition of $D(y(r, t))$
- $\text{Inv}(A)$ = set of all A -invariant subspaces
- i = positive integer
- j = positive integer
- k = positive integer
- m = positive integer
- \mathcal{M} = subspace of n -dimensional space
- \mathcal{M}_G = subspace spanned by the row vectors of M_G
- n = dimension of vector y
- \hat{n} = dimension of vector \hat{y}

- \mathcal{R} = field of real number
 \mathcal{R}^n = n -dimensional real space
 s_k = the rank of A_k or equal to $n - 1$
 t = time
 V = interior of the catalyst particle
 ∂V = boundary of V
 y_k = k th element of vector y

Vectors and Matrices

Capital letters represent matrices; bold-face lower case letters represent vectors.

- A = constant matrix
 A_0 = constant matrix
 A_k = constant matrix
 B_0 = defined as $A_0 D^{-1}$
 B_k = defined as $A_k D^{-1}$
 D = effective diffusivity matrix
 D_i = constant basis matrix of $D(\mathbf{r})$, $D(\mathbf{y}(\mathbf{r}))$ or $D(\mathbf{y}(\mathbf{r}, t))$
 $D(\mathbf{r})$ = effective diffusivity matrix, which is a function of position
 $D(\mathbf{y}(\mathbf{r}))$ = effective diffusivity matrix, which is a function of concentrations
 $D(\mathbf{y}(\mathbf{r}, t))$ = effective diffusivity matrix, which is a function of concentrations
 \hat{D} = nonsingular matrix
 $\hat{D}(\mathbf{r})$ = nonsingular matrix
 $\hat{D}(\mathbf{y}(\mathbf{r}))$ = nonsingular matrix
 $\hat{D}(\mathbf{y}(\mathbf{r}, t))$ = nonsingular matrix
 $\mathbf{f}(\mathbf{y}(\mathbf{r}))$ = n -dimensional function vector
 $\mathbf{f}(\mathbf{y}(\mathbf{r}, t))$ = n -dimensional function vector

$\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}))$	= \hat{n} -dimensional function vector
$\hat{\mathbf{f}}(\hat{\mathbf{y}}(\mathbf{r}, t))$	= \hat{n} -dimensional function vector
$\mathbf{g}(\mathbf{z}(\mathbf{r}))$	= n -dimensional function vector
I	= identity matrix
$J(\mathbf{y}(\mathbf{r}))$	= Jacobian matrix of $\mathbf{f}(\mathbf{y}(\mathbf{r}))$
$J(\mathbf{y}(\mathbf{r}, t))$	= Jacobian matrix of $\mathbf{f}(\mathbf{y}, t)$
$J(\mathbf{z}(\mathbf{r}))$	= Jacobian matrix of $\mathbf{g}(\mathbf{z}(\mathbf{r}))$
K	= rate constant matrix
L	= linear operator
\hat{L}	= linear operator
$\mathbf{m}_{(i)}$	= i th row vector of M
M	= lumping matrix
\bar{M}	= generalized inverse of M satisfying $M\bar{M} = I_n$
M_G	= given submatrix of M .
Q	= $\hat{n} \times \hat{n}$ constant matrix
$Q(\mathbf{y}(\mathbf{r}))$	= $\hat{n} \times \hat{n}$ matrix
$Q(G)_{(ki)}^T$	= orthonormal matrix representations of $\text{Im}(M_G(A_k^T)^i)^T$
\mathbf{r}	= position vector
\mathbf{R}	= position vector on the boundary
R	= eigenvector matrix of Y
\mathbf{x}	= n -dimensional vector
X	= eigenvector matrix or an $n \times (n - \hat{n})$ matrix
$\mathbf{y}(\mathbf{r})$	= n -dimensional variable vector
$\mathbf{y}(\mathbf{r}, t)$	= n -dimensional variable vector
$\hat{\mathbf{y}}(\mathbf{r})$	= \hat{n} -dimensional variable vector
$\hat{\mathbf{y}}(\mathbf{r}, t)$	= \hat{n} -dimensional variable vector
Y	= defined as $\sum_{k=1}^m \sum_{i=0}^{k-1} Q(G)_{(ki)}^T Q(G)_{(ki)}$

$z(r)$ = n -dimensional variable vector

Greek Letters

α_i = real number

β_i = real number

Λ = diagonal eigenvalue matrix with λ_i as its i th diagonal element

Symbols

$\hat{}$ = any property related to the lumped system

0 = null vector

0 = null matrix

REFERENCES

Ben-Israel, A. and Greville, T.N.E., 1974, Generalized Inverse: Theory and Applications, John Wiley & Sons, Inc., New York.

Gohberg, I., Lancaster, P. and Rodman, L., 1986, Invariant Subspaces of Matrices with Applications, John Wiley & Sons, Inc., New York.

Lang, S., 1986, Introduction to Linear Algebra, 2nd edition, Springer-Verlag, New York.

Li, G., 1984, A lumping analysis in mono- or/and bimolecular reaction systems, Chem. Eng. Sci., **39**, 1261-1270.

Li, G. and Rabitz, H., 1989, A general analysis of exact lumping in chemical kinetics, Chem. Eng. Sci., **44**, 1413-1430.

Li, G. and Rabitz, H., 1990a, A general analysis of approximate lumping in chemical kinetics, Chem. Eng. Sci., 45, 977-1002.

Li, G. and Rabitz, H., 1990b, New approaches to determination of constrained lumping schemes for a reaction system in the whole composition space, Chem. Eng. Sci., 45, in print.

Wei, J., 1962, Intraparticle diffusion effects in complex systems of first order reactions, J. of Catalysis, 1, 526-546.

Wei, J. and Kuo, J.C.W., 1969, A lumping analysis in monomolecular reaction systems, Ind. Eng. Chem. Fundamentals, 8, 114-133.

APPENDIX

We will prove that when the subspace \mathcal{M} spanned by the row vectors of the lumping matrix M is $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ -invariant, then this condition is sufficient for exact lumping of a reaction system coupled with diffusion under steady-state conditions.

Suppose $J^T(\mathbf{y}(\mathbf{r}))D^{-1}$ has a nontrivial fixed \hat{n} -dimensional invariant subspace \mathcal{M} with the $(n \times \hat{n})$ -matrix representation M^T . Its orthogonal complement is \mathcal{N} in the n -dimensional space with the $(n \times (n - \hat{n}))$ -matrix representation X . In order to simplify the discussion we choose two sets of orthonormal bases for \mathcal{M} and \mathcal{N} , i.e.,

$$MM^T = I_{\hat{n}}, \quad (A.1)$$

$$X^T X = I_{n-\hat{n}}. \quad (A.2)$$

Therefore, the matrix $(X|M^T)$ is an orthogonal one and its inverse is just the transpose of itself: $\begin{pmatrix} X^T \\ M \end{pmatrix}$. Then we have

$$\begin{pmatrix} X^T \\ M \end{pmatrix} (X|M^T) = (X|M^T) \begin{pmatrix} X^T \\ M \end{pmatrix} = I_n. \quad (A.3)$$

For the following nonsingular linear transformation

$$\mathbf{z}(\mathbf{r}) = \begin{pmatrix} X^T \\ M \end{pmatrix} \mathbf{y}(\mathbf{r}), \quad (A.4)$$

we have the inverse transformation

$$\mathbf{y}(\mathbf{r}) = (X|M^T) \mathbf{z}(\mathbf{r}), \quad (A.5)$$

and

$$\begin{aligned} \nabla^2 \mathbf{z}(\mathbf{r}) &= \begin{pmatrix} X^T \\ M \end{pmatrix} \nabla^2 \mathbf{y}(\mathbf{r}) \\ &= - \begin{pmatrix} X^T \\ M \end{pmatrix} D^{-1} \mathbf{f}(\mathbf{y}(\mathbf{r})) \\ &= - \begin{pmatrix} X^T \\ M \end{pmatrix} D^{-1} \mathbf{f}((X|M^T) \mathbf{z}(\mathbf{r})) \\ &= \mathbf{g}(\mathbf{z}(\mathbf{r})). \end{aligned} \quad (A.6)$$

The corresponding Jacobian matrix of $g(z(r))$ is

$$\begin{aligned}
J(z(r)) &= -\partial \left(\begin{pmatrix} X^T \\ M \end{pmatrix} D^{-1} f((X|M^T)z(r)) \right) / \partial z(r) \\
&= - \begin{pmatrix} X^T \\ M \end{pmatrix} D^{-1} \frac{\partial}{\partial y(r)} f(y(r)) \frac{\partial y(r)}{\partial z(r)} \\
&= - \begin{pmatrix} X^T \\ M \end{pmatrix} D^{-1} J(y(r)) (X|M^T) \\
&= - \begin{pmatrix} X^T D^{-1} J(y(r)) X & X^T D^{-1} J(y(r)) M^T \\ M D^{-1} J(y(r)) X & M D^{-1} J(y(r)) M^T \end{pmatrix}. \quad (A.7)
\end{aligned}$$

When the subspace \mathcal{M} spanned by the row vectors of M is a fixed invariant one of $J^T(y(r))D^{-1}$ for all values of $y(r)$, i.e., a left fixed invariant subspace of $D^{-1}J(y(r))$ for all values of $y(r)$, we have

$$M D^{-1} J(y(r)) X = Q(y(r)) M X = 0, \quad (A.8)$$

where $Q(y(r))$ is an $\hat{n} \times \hat{n}$ matrix and then Equation A.7 becomes

$$J(z(r)) = - \begin{pmatrix} X^T D^{-1} J(y(r)) X & X^T D^{-1} J(y(r)) M^T \\ 0 & M D^{-1} J(y(r)) M^T \end{pmatrix}. \quad (A.9)$$

Since the transformation in Equation A.4 is nonsingular, all values of $y(r)$ means all values of $z(r)$. Therefore from Equation A.9 we have

$$\partial g_i(z(r)) / \partial z_j(r) = 0. \quad (A.10)$$

$$(i = n - \hat{n} + 1, n - \hat{n} + 2, \dots, n; j = 1, 2, \dots, n - \hat{n}) \quad \forall z(r) \in R^n$$

Equation A.10 shows that $g_i(z(r)) (i = n - \hat{n} + 1, n - \hat{n} + 2, \dots, n)$ do not contain the first $n - \hat{n}$ elements $z_j(r) (j = 1, 2, \dots, n - \hat{n})$. Therefore, the last \hat{n} equations in Equation A.6 compose an exactly lumped model.

Now we will demonstrate that this lumped model can be represented as

$$\nabla^2 \hat{y}(r) = -M D^{-1} f(\bar{M} \hat{y}(r)). \quad (A.11)$$

Let

$$\hat{\mathbf{y}}(\mathbf{r}) = M\mathbf{y}(\mathbf{r}). \quad (\text{A.12})$$

From Equation A.6 one has

$$\nabla^2 \hat{\mathbf{y}}(\mathbf{r}) = -MD^{-1}\mathbf{f}((X|M^T)\mathbf{z}(\mathbf{r})). \quad (\text{A.13})$$

Considering that these equations do not contain $\mathbf{z}_j(\mathbf{r}) (j = 1, 2, \dots, n - \hat{n})$, Equation A.13 is equivalent to

$$\begin{aligned} \nabla^2 \hat{\mathbf{y}}(\mathbf{r}) &= -MD^{-1}\mathbf{f}((0|M^T)\mathbf{z}(\mathbf{r})) \\ &= -MD^{-1}\mathbf{f}(M^T \hat{\mathbf{y}}(\mathbf{r})). \end{aligned} \quad (\text{A.14})$$

Multiplying Equation 4 in Section IIA from the left by M and comparing the resultant equations with Equation A.14 yields

$$\begin{aligned} MD^{-1}\mathbf{f}(\mathbf{y}(\mathbf{r})) &= MD^{-1}\mathbf{f}(M^T \hat{\mathbf{y}}(\mathbf{r})) \\ &= MD^{-1}\mathbf{f}(M^T M\mathbf{y}(\mathbf{r})). \end{aligned} \quad (\text{A.15})$$

This holds for any values of $\mathbf{y}(\mathbf{r})$. Therefore, letting $\mathbf{y}(\mathbf{r})$ take the value $\bar{M}\hat{\mathbf{y}}(\mathbf{r})$, we have

$$\begin{aligned} MD^{-1}\mathbf{f}(\bar{M}\hat{\mathbf{y}}(\mathbf{r})) &= MD^{-1}\mathbf{f}(M^T M\bar{M}\hat{\mathbf{y}}(\mathbf{r})) \\ &= MD^{-1}\mathbf{f}(M^T \hat{\mathbf{y}}(\mathbf{r})). \end{aligned} \quad (\text{A.16})$$

Substituting Equation A.16 into Equation A.14 gives Equation A.11.

Appendix K

11. Lie Algebraic Factorization of Multivariable Evolution Operators:
Convergence Theorems for the Canonical Case, M. Demiralp and H. Rabitz,
Int. J. of Eng. Sci., in press.

**LIE ALGEBRAIC FACTORIZATION OF MULTIVARIABLE
EVOLUTION OPERATORS: CONVERGENCE THEOREMS
FOR THE CANONICAL CASE***

Metin Demiralp and Herschel Rabitz**

Princeton University, Department of Chemistry
Princeton, N.J. 08544-1009, USA

* Supported by NATO via RG.86/0123, the Office of Naval Research and the Air Force Office of Scientific Research.

** Permanent Address: İstanbul Technical University, Faculty of Sciences and Letters, Engineering Sciences Department, Ayazağa Campus, Maslak, 80626 - İstanbul, TURKEY

ABSTRACT

This work is devoted to establishing the convergence theorems for the canonical case of the Lie algebraic factorization of multivariable evolution operators. The definition and various properties of $\bar{\xi}$ -approximants are given in a companion paper. The theorems presented in this paper give some sufficient conditions for the convergence of the $\bar{\xi}$ -approximant sequences. Proofs are given for a specific region of the variables space appearing in the Lie operator and the theorems are useful for many practical applications.

1. INTRODUCTION

In a companion paper [1], we have given certain transformations which are based on a space extension concept, to put the Lie evolution operator into a new form potentially amenable to practical computation. The latter paper reduced the general case to a canonical problem for the Lie algebraic factorization of multivariable evolution operators. In particular, we reduced the structure of the descriptive functions f_1, \dots, f_N in $\mathbf{f} \cdot \nabla$ (Lie-operator) to a quadratic one by assuming a closedness condition on the components f_1, \dots, f_N under the action of ∇ via a space extension technique. This extension, (it may be a contraction in certain special cases) brings us to the canonical case where the linear response of the system is characterized by $\lambda \mathbf{I}$ (\mathbf{I} is the unit matrix). The importance of the canonical case lies in the fact that the σ -coefficients which generate the $\bar{\xi}$ -approximants [1] can be evaluated via finite step algorithms in an analytical way.

The linear response matrix which generates the linear terms of the extended descriptive functions affects the convergence properties of $\bar{\xi}$ -approximants, and it is important to manipulate its structure via the available parameters, $(\lambda, \nu_1, \nu_2, \dots, \nu_N)$, which enter the space extension to change the factorization problem into a canonical one (See ref.[1] for details). Since all these parameters give a certain flexibility to change the behaviour of the linear response matrix, we are able to obtain the most appropriate linear response matrix for our purposes.

We use an N -parameter unitary transformation when we rotate the axes of the space of the variables of the Lie operator to get a factorization point placed on x_1 -axis, $[1, 0, \dots, 0]$. Hence, depending on these N -parameters, the $\bar{\xi}$ -approximants of the factorization can have different structures. As we recall, any component of the vector resulting from the action of the evolution operator on the position vector can be expressed as a linear combination of $(N+1)$ different $\bar{\xi}$ -approximants. Therefore, we have to use $(N+1)$ different $\bar{\xi}$ -approximant sequences for a real multivariable factorization scheme. This is the main difference between the multivariable and one-variable factorization schemes.[1-3]. However, a most important result is the lack of coupling among these different sequences. In other words, each of this $(N+1)$ different $\bar{\xi}$ -approximant sequences, can be constructed through first order recursions between $\bar{\xi}_{n+1}$ and $\bar{\xi}_n$ without regard to the other sequences. The arbitrariness

arising in the choice of the ν -parameters gives the necessary flexibility to significantly adjust the convergence of the $\bar{\xi}$ -approximant sequences.

The next section will include some preliminary discussion about the convergence properties of the $\bar{\xi}$ -approximant sequences. Third section is devoted to the detailed convergence analysis. Certain lemmas and theorems will be given with their proofs. The fourth section will present the concluding remarks.

2. THE SINGULARITIES OF THE $\bar{\xi}$ -APPROXIMANTS

In the canonical case, the multivariable evolution operator to be factorized has the following form

$$Q = \exp\{f(z) \cdot \nabla\} \quad (2.1)$$

where $f(z)$ is a given specific $(N + 1)$ -dimensional vector function which defines the Lie operator under consideration. The linear response matrix of the system is assumed to be proportional to the unit matrix. The proportionality constant is denoted by λ and is called the "Characteristic Mode". The action of Q on a component of z , say z_j , is approximated by a linear combination of $(N + 1)$ -different $\bar{\xi}$ -approximants,

$$Qz_j \approx \sum_{m=1}^{N+1} c_m^{(j)} \bar{\xi}_{n,m} \quad (2.2)$$

such that

$$\bar{\xi}_{n+1,m} = \frac{\bar{\xi}_{n,m}}{\sqrt[n]{1 - n\sigma_{n+1}^{(m)} \bar{\xi}_{n,m}}} \quad (2.3)$$

where n is the recursion index and the $\sigma_n^{(m)}$ and $c_m^{(j)}$ coefficients depend on, $(\nu_1, \dots, \nu_{N-1})$, the arbitrary parameters of the rotation which is used to bring the factorization point onto the x_1 -axis and the $\bar{\xi}$ -approximants implicitly depend on j , not indicated for notational reasons. The initial element, $\bar{\xi}_1$ of the $\bar{\xi}$ -approximant sequence can be given as follows

$$\bar{\xi}_{1,m} = e^{\lambda_m t} \quad (2.4)$$

where λ_m stands for one of the characteristic modes. Although there is only one characteristic modal value in the descriptive functions of the system under consideration, it

depends on the convergence control parameters, $\nu_1, \nu_2, \dots, \nu_N$ and may take different values for each different selection of the ν -values. Since we use $(N + 1)$ -different set of ν -values when we generate the action of Q on each separate coordinate, there will be a possibility of producing $(N + 1)$ characteristic modal values, $\lambda_1, \lambda_2, \dots, \lambda_{N+1}$. These values may not actually represent the true characteristic modes of the system due to the fact that the evaluation of the characteristic modes of a given system may become quite difficult when we deal with nonlinear systems. However, they must satisfy certain global features for the sake of numerical convergence. For example, λ -values must have non-zero imaginary parts when we deal with a pure oscillatory system. There is no restriction on the ν -parameters, however we can specify them in a way such that the convergence rate of $\bar{\xi}$ -approximant sequences is maximal.

Now, let us consider the recursion given by the Eq(2.3). Recalling that [1]

$$\sigma_2^{(m)} = b_{111}^{(m)} \frac{1 - e^{-\lambda t}}{\lambda} \quad (2.5)$$

we can express $\bar{\xi}_2$ as follows

$$\bar{\xi}_2 = \sum_{m=1}^{N+1} c_m^{(j)} e^{\lambda_m t} \frac{1}{1 - b_{111}^{(m)} \frac{e^{\lambda_m t} - 1}{\lambda}} \quad (2.6)$$

where the superscript, m , characterizes the ν -dependence of the corresponding entity. This formula reveals the singular structure of the $\bar{\xi}$ -approximants and gives clues about the types of the singularities which may appear in anyone of the elements of the $\bar{\xi}$ -approximant sequences. Before proceeding further, we confine ourselves to this quite simple case.

The right hand side of Eq.(2.6) has certain poles to the $b_{111}^{(m)}$ -parameters. This can be more clearly explained as follows: If we consider the right hand side of the Eq(2.6) as a mapping whose domain is the cartesian sum of $b_{111}^{(m)}$ -complex planes, ($m = 1, 2, \dots, N + 1$) then, every individual $b_{111}^{(m)}$ -complex plane has a pole varying with time. At the beginning of the evolution ($t = 0$) these poles are gathered at infinity, and their location moves toward the origin of the corresponding $b_{111}^{(m)}$ -complex plane. This structure is reminiscent of the Padé approximants. Since we can increase the number of variables by taking second degree terms in z as new variables, this does not destroy the quadratic structure of the system. We can even recover the canonical structure by extending the space via a new variable which

is simply equal to one. The consecutive use of these transformations causes an increase in the number of the poles of the right hand side of the Eq(2.6). Hence the second order $\bar{\xi}$ -approximants have a nature which is quite similar to the Padé approximants. However the following distinctions exist.

i) Padé approximants have a single complex plane as a domain, but the $\bar{\xi}_2$ -approximant's domain is composed of $(N + 1)$ separate complex planes.

ii) The poles of the Padé approximants are motionless unless the function which generates the Padé approximants has coefficients varying with respect to time (or with respect to a corresponding parameter).

iii) $\bar{\xi}_2$ -approximants can only be related to a special sequence of Padé approximants (placed into the lower diagonal adjacent to the main diagonal) and the style of increase in the order is different for both approximants. Indeed, Padé approximant's order is increased by one, however, the increase in the order of the $\bar{\xi}_2$ -approximant is determined by the number of second degree terms used in the space extension mentioned above.

Similar comparisons can also be made for other $\bar{\xi}$ -approximants and certain connections can be established between Hermite-Padé approximants and $\bar{\xi}_n$ -approximants. For higher n values branch points appear in the structure of $\bar{\xi}_n$ and the domain of the transformation characterized by $\bar{\xi}_n$ is again the cartesian product of $b_{111}^{(m)}$ -complex planes. However, each of these planes must be appropriately cut to take care of the branch points. The shapes and locations of these cuts vary with time due to the time-dependence of the branch points. As long as we consider finite values of n , these branch points are algebraic, however this algebraic structure approaches a logarithmic limit one when n goes to infinity. Similar behaviour can be observed in the Hermite-Padé approximants but the spirit of the construction of both approximants are quite different because of their typically distinct purposes. Since this issue is beyond the scope of this work we shall not get into further details of this topic. However, we can say comment on the singular behaviour of the $\bar{\xi}$ -approximants as follows:

i) Each singularity of the $\bar{\xi}$ -approximants belongs to a specified $b_{111}^{(m)}$ -complex plane and it always remains in the same plane during the evolution.

ii) Whether poles or branch points, all singularities are gathered at infinity in the composite space of $b_{111}^{(m)}$ -complex planes at the beginning of the evolution. Each singularity

moves along a trajectory in its corresponding $b_{111}^{(m)}$ -complex plane as time evolves and may or may not reach the origin when t tends to infinity.

iii) If none of the singularities reaches to the origin, then every $b_{111}^{(m)}$ -complex plane has a "Clean Region", into which a singularity trajectory never enters during the evolution.

iv) We call the union of these clean regions as the "Main Clean Region" of the system. Here we use the word "system" to characterize a collection of variables; it is not meant in a system-theoretical meaning.

Therefore every system has a main clean region during a finite evolution $t \in [0, T]$ for an appropriate value of T . Depending on T we use the following designations:

- a) If $T = \infty$, then the system is "Global Normal".
- b) If T has a finite non-zero value, then the system is "Temporary Normal".
- c) If $T = 0$, then the system is "Abnormal".

This terminology follows the earlier work [2,3] and will be utilized below.

3. CONVERGENCE THEOREMS

Now, we are ready to proceed to prove certain convergence theorems. For this purpose, we consider the following simplest one of the general multivariable factorization problems, the canonical factorization problem

$$\{Qz_1\}_{z=e_1} = \{e^{if(z) \cdot \nabla}\}_{z=e_1} \quad (3.1)$$

where z and ∇ are the position vector and gradient operator in an N -dimensional complex Euclidean space. The vector function, $f(z)$, is given as below

$$f_i(z) = \lambda z_i + \sum_{j=1}^N \sum_{k=1}^N b_{ijk} z_j z_k \quad i = 1, \dots, N \quad (3.2)$$

Here and in the coming sections, e_1 stands for the unit cartesian vector $[1, 0, \dots, 0]$.

The vector e_1 is apparently an eigenvector of the linear response matrix λI . The unit matrix structure in the linear response term is due to the canonical structure of the problem. As shown in the companion paper [1], the assumption of a canonical structure of the problem does not cause any loss of generality because we can always convert a quadratic structure to a canonical one by means of a simple space extension. Indeed, almost

every factorization problem can be brought into the canonical one unless the structure of descriptive functions prevent us to find a proper space extension to this end. The expense of this procedure is an increase of the number of independent variables. Since these transformations involve a finite number of steps, the theorems proved for the rather simple factorization problem, also remain valid for the original factorization problem before the space extension transformation.

The factorization of the evolution operator given by the Eq(3.1) can be expressed as follows

$$\{Qz_1\}_{\mathbf{z}=\mathbf{e}_1} = \left\{ e^{t\mathbf{s}^T \nabla} \left\{ \prod_{j=1}^{\infty} e^{\mu_j z_1 \frac{\partial}{\partial z_j}} \right\} z_1 \right\}_{\mathbf{z}=\mathbf{e}_1} \quad (3.3)$$

where μ_j depends on z_2, \dots, z_N and t . The non-existence of terms including operators corresponding to differentiation with respect to the other coordinates z_2, \dots, z_N is due to the selection of a special ordering for the simple evolution operators such that their effects on z_1 are nothing except multiplication by unity. Furthermore, the dependence of μ_j -functions on z_2, \dots, z_N can be removed since the factorization can be evaluated at a special point where z_2, \dots, z_N vanish. Hence we can simply write

$$\{Qz_1\}_{\mathbf{z}=\mathbf{e}_1} = \left\{ \left\{ \prod_{j=1}^{\infty} e^{\sigma_j z_1 \frac{\partial}{\partial z_j}} \right\} z_1 \right\}_{z_1=1} \quad (3.4)$$

where

$$\sigma_j(t) = \mu_j(0, 0, \dots, t) + \lambda \delta_{j1} \quad j = 1, \dots, \infty \quad (3.5)$$

and δ_{jk} denotes the Kronecker's delta. Here, we have used the fact that μ_0 and μ_1 vanish when all the z_j -variables except z_1 tend to zero.

As an approximation, we define the ξ -approximants as follows

$$\{Qz_1\}_{\mathbf{z}=\mathbf{e}_1} \approx \left\{ \left\{ \prod_{j=1}^n e^{\sigma_j z_1 \frac{\partial}{\partial z_j}} \right\} z_1 \right\}_{z_1=1} \equiv e^{\lambda t} \xi_n \quad (3.6)$$

By using properties of Lie operators we can prove that these approximants satisfy the following recursion

$$\xi_1 = 1 \quad (3.7a)$$

$$\xi_{n+1} = \frac{\bar{\xi}_n}{\sqrt[3]{1 - n\sigma_{n+1} e^{n\lambda t} \xi_n^n}} \quad n = 1, 2, \dots \quad (3.7b)$$

Obviously, this recursion is a mapping from the complex plane of ξ_n to the complex plane of ξ_{n+1} , and the ξ_n -plane must be properly cut to take care of branch points. The derivation of the recursion for the ξ -approximants is based on certain properties of Lie evolution operators. These properties are derived via Taylor series expansion, so one can expect that their validities are limited by the convergence domain of Taylor series, and this means that the validity of the recursion relation of ξ -approximants is also limited by an appropriate contour surrounding the origin of the ξ_n -complex plane. However, by analytic continuation of the Taylor series outside their convergence domains, the same type of the generalization of the recursion of the ξ -approximants to outside their convergence domain defined by the contours in ξ_n -complex plane should also be possible. This means that the recursion between ξ_n and ξ_{n+1} remains valid for the entire complex plane of ξ_n except the branch cuts. So, we can interpret the recursion between two consecutive ξ -approximants as follows:

i) Each ξ -approximant corresponds to a point in its own complex plane, and there are an infinite number of complex planes. Since the n -th complex plane is the domain of the mapping between ξ_n and ξ_{n+1} , it is composed of n numbers of Riemann sheets due to the n -th order algebraic branch point appearing in the recursion between ξ_n and ξ_{n+1} .

ii) Our factorization point is to be considered as a point in the complex plane of ξ_1 . Since there is no branch point in the mapping between ξ_1 and ξ_2 , the only singularity is a pole accordingly moving as time evolves.

iii) The ξ_n -complex plane is related to the ξ_1 -complex plane through n numbers of consecutive mappings. Hence, as being the domain of this composite mapping, the ξ_1 -complex plane must have a structure such that it can take care of all branch points appearing in the intermediate stages of this mapping. Obviously this structure changes depending on n . Since our essential goal is to characterize the evolution under ξ_∞ , the most important form of the ξ_1 -complex plane is its structure appearing when n increases to infinity. In this case, there appear an infinite number of moving branch point trajectories and the behavior of these trajectories like their locations etc., completely determines the nature of the evolution. However, instead of the considering the entire composite mapping, the use of individual mapping is easy and it facilitates a better understanding of the character of the evolution.

Since our present purpose is to establish the proofs for the convergence of the ξ -approximant sequences, not for the entire complex plane of ξ_1 but for certain clean regions,

we shall leave further investigation of entire plane convergence of ξ -approximants to a future work.

Although the convergence properties of the recursion between ξ_n and ξ_{n+1} was shown in one of our previous works [2,3], we briefly summarize it here to facilitate an understanding of the proofs of the theorems of present work. Now, as we can see, the Eqs(3.7a) and (3.7b) permit us to write

$$\xi_n = \frac{1}{\Delta_n(\xi_1, t)} \quad n = 1, 2, \dots \quad (3.8)$$

and this results in the following recursion between Δ_n and Δ_{n+1}

$$\Delta_{n+1} = \sqrt[n]{\Delta_n^n - n\sigma_{n+1}\xi_1^n e^{n\lambda t}} \quad \Delta_1 = 1 \quad (3.9)$$

where ξ_1 is used to specify the ξ_1 -dependence of relevant entities and its value will be equated to 1 later. Let us, now, consider a majorant function, $D(\xi_1, t)$ which converges in a certain region of the ξ_1 -complex plane, the time-dependent convergence radius of which is denoted by $\zeta_n(t)$ such that $D(\xi_1, t)$ remains greater than 1 and also greater than Δ_n for this region. By appropriately increasing the value of the right hand side of the first one of the Eqs(3.9) and using D instead of Δ we can arrive at the following recursion between D_n and D_{n+1}

$$D_{n+1}(\xi_1, t) = D_n(\xi_1, t) \{1 + (n+1)|\sigma_{n+1}||\xi_1|^n e^{n\Re(\lambda)t}\} \quad (3.10)$$

The consecutive use of this equation from a prescribed value of n , say N , to infinity enables us to write

$$D_\infty(\xi_1, t) = D_N(\xi_1, t) \prod_{j=1}^{\infty} \{1 + (N+j)|\sigma_{N+j}||\xi_1|^{N+j} e^{(N+j)\Re(\lambda)t}\} \quad (3.11)$$

The condition for the convergence of the infinite product appearing in the last equation is related to the convergence of the following infinite sum

$$d_N(\xi_1, t) = \sum_{j=1}^{\infty} (N+j)|\sigma_{N+j}||\xi_1|^{N+j} e^{(N+j)\Re(\lambda)t} \quad (3.12)$$

If this sum converges for certain ξ_1 , t and sufficiently large N values and it tends to zero as N increases unboundedly, then the infinite product in the Eq(3.11) also converges for same ξ_1 and t values.

Since the σ -coefficients depend on time, the convergence radius of the infinite product in the ξ_1 -complex plane varies with time. If we denote this convergence radius by $\zeta(t)$ and its minimum value by $\zeta_{\min}(t)$ for $t \in [0, \infty)$ then the following circumstances may occur:

i) $\zeta_{\min}(t)$ is greater than zero, then, there is, at least, one "Non-empty Clean Region" around the origin of the ξ_1 -complex plane.

ii) $\zeta_{\min}(t)$ equals to zero, then, there is no region which remains clean during the entire evolution. However, even in this case, one can find a temporary minimum convergence radius, $\zeta_{\min}(T)$ such that it does not vanish for a finite time period $t \in [0, T]$; then, there is, at least, one "Non-empty Temporary Clean Region" around the origin of the ξ_1 -complex plane.

iii) If the temporary minimum convergence radius function, $\zeta_{\min}(T)$ vanishes for any finite time period, then there is no "Temporary or Permanent Clean Region" around the origin of ξ_1 -complex plane. The system under consideration is, then, an "Abnormal System".

So, we have proved the following theorem.

THEOREM 1:

If the following infinite sum

$$d(\xi_1, t) = \sum_{j=1}^{\infty} (j+1) |\sigma_{j+1}| |\xi_1|^j e^{j\Re(\lambda)t} \quad (3.13)$$

converges in a circle around the origin of the ξ_1 -complex plane, the radius of which is $\zeta(t)$, then the following statements are valid:

i) If $\zeta(t) \geq \zeta_{\min}(t) > 0$ for $t \in [0, \infty)$, then, the system is "Global Normal".

ii) If $\zeta(t) \geq \zeta_{\min}(T) > 0$ for $t \in [0, T)$ with $T > 0$, then, the system is, at least, "Temporary Normal".

As a corollary we can say that if the first condition of Theorem 1 holds then the sequence of ξ -approximants converges for all ξ_1 and t values in the regions defined as $|\xi_1| < \zeta_{\min}(t)$ and $t \in [0, \infty)$ respectively, and they have a permanent main clean region which is not empty with respect to an appropriately defined measure.

Let us, now, consider the following linear form in z_1, z_2, \dots, z_N

$$h_1 = \sum_{j=1}^N c_j z_j \quad (3.14)$$

where the c -coefficients are different than the formerly employed ones. A brief look at the structure of h_1 shows that

$$|h_1| < \sqrt{\sum_{j=1}^N |c_j|^2} \tau \quad (3.15)$$

where τ denotes the hyperradial variable in N -dimensional complex Euclidean space of z -variables as below

$$\tau = \sqrt{\sum_{j=1}^N |z_j|^2} \quad (3.16)$$

We can also write the following inequality for the derivatives of h_1 in the same manner

$$\left| \frac{\partial h_1}{\partial z_k} \right| < \sqrt{\sum_{j=1}^N |c_j|^2} \quad k = 1, \dots, N \quad (3.17)$$

A simple but somewhat detailed analysis shows that the following inequalities hold for these quadratic forms (second degree forms)

$$|h_2| < \sqrt{\sum_{j=1}^N \sum_{k=1}^N |c_{jk}|^2} \tau^2 \quad (3.18)$$

and

$$\left| \frac{\partial h_2}{\partial z_k} \right| < \sqrt{\sum_{j=1}^N \sum_{k=1}^N |c_{jk}|^2} \tau \quad k = 1, \dots, N \quad (3.19)$$

These results can be easily generalized to the n -th order forms via mathematical induction.

To this end, we can assume that the following formulas are valid

$$h_n = \sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_n=1}^N c_{j_1 j_2 \dots j_n} z_{j_1} z_{j_2} \dots z_{j_n} \quad (3.20)$$

$$|h_n| < \sqrt{\sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_n=1}^N |c_{j_1 j_2 \dots j_n}|^2} \tau^n \quad (3.21)$$

$$\left| \frac{\partial h_n}{\partial z_k} \right| < n \sqrt{\sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_n=1}^N |c_{j_1 j_2 \dots j_n}|^2} \tau^{n-1} \quad k = 1, \dots, N \quad (3.22)$$

then we can express h_{n+1} in terms of certain n -th order forms as follows

$$h_{n+1} = \sum_{j=1}^N h_n^{(j)} z_j \quad (3.23)$$

$$\frac{\partial h_n}{\partial z_k} = h_n^{(k)} + \sum_{j=1}^N \frac{\partial h_n^{(j)}}{\partial z_k} z_j \quad (3.24)$$

where

$$h_n^{(j)} = \sum_{j_2=1}^N \sum_{j_3=1}^N \dots \sum_{j_n=1}^N c_{j j_2 j_3 \dots j_n} z_{j_2} z_{j_3} \dots z_{j_n} \quad (3.25)$$

By using the Cauchy-Schwartz inequality for scalar products we can obtain the following inequalities

$$|h_{n+1}| < \sqrt{\sum_{j=1}^N |h_n^{(j)}|^2} r \quad (3.26)$$

$$\left| \frac{\partial h_{n+1}}{\partial z_k} \right| < |h_n^{(k)}| + \sqrt{\sum_{j=1}^N \left| \frac{\partial h_n^{(j)}}{\partial z_k} \right|^2} r \quad (3.27)$$

If we compare the Eq(3.26) with the Eq(3.21) we can conclude that the Eq(3.21) remains valid when n is replaced with $(n+1)$, so its validity for all positive integer values of n has been proved. However, the proof of Eq(3.22) necessitates a little more detailed analysis. To this end, we can increase the value of the first term of the right hand side of the Eq(3.27) by replacing it with the square root of the sum over the squares of its values for $k \in [1, \dots, N]$. Then, we are able to show that Eq(3.22) remains valid for all positive integer values of n .

Therefore we can easily arrive at the following lemma via appropriate intermediate steps

LEMMA 1:

Consider a multivariable function, $H(z_1, \dots, z_N)$ which can be expanded into a series of homogeneous multinomials as follows

$$H(z_1, \dots, z_N) = \sum_{n=0}^{\infty} h_n(z_1, \dots, z_N) \quad (3.28)$$

where

$$h_n(z_1, \dots, z_N) = \sum_{j_1=1}^N \sum_{j_2=1}^N \dots \sum_{j_n=1}^N c_{j_1 j_2 \dots j_n} z_{j_1} z_{j_2} \dots z_{j_n} \quad (3.29)$$

This function and its first order partial derivatives with respect to z -variables are majorized by the following functions of hyperradius given by the Eq(3.16)

$$|H(z_1, \dots, z_N)| < H_M(r) \quad (3.30)$$

$$\left| \frac{\partial H}{\partial z_k} \right| < \frac{\partial H_M}{\partial r} \quad (3.31)$$

where

$$H_M(r) = \sum_{k=0}^{\infty} H^{(k)} r^k \quad (3.32)$$

and

$$H^{(k)} = \sqrt{\sum_{j=1}^N \dots \sum_{k=1}^N |c_{j_1 \dots j_k}|^2} \quad (3.33)$$

Now we are sufficiently equipped for the derivation of a majorant function to use in the convergence proof of the ξ -approximants. To this end we can consider to seek bounds for the σ -coefficients. Since the σ -parameters are only special values of the corresponding μ -functions, we are going to deal with μ -functions instead of σ -functions. So, we rewrite the equations for the μ -functions, which are formerly given in the companion paper [1] of this work as follows

$$\frac{\partial \mu_0}{\partial t} = \{f_1^{(0)}(t, z - \mu_0 e_1) + \sum_{k=2}^N f_k^{(0)}(t, z - \mu_0 e_1) \frac{\partial \mu_0}{\partial z_k}\}_{z_1=0} \quad (3.34)$$

$$f_j^{(0)}(t, z) = e^{\lambda t} \left\{ \sum_{k=1}^N \sum_{l=1}^N b_{jkl} z_k z_l \right\} \quad (3.35)$$

$$f_k^{(m+1)}(t, z) = [f_k^{(m)}(t, z)]^* \quad (3.36)$$

$$f_1^{(m+1)}(t, z) = \frac{F_m(t, z) - \{F_m(t, z)\}_{z_1=0}}{x_1} \quad m \geq 0 \quad (3.37)$$

$$F_m(t, x) = [f_1^{(m)}(t, z) + \sum_{k=2}^N f_k(t, z) \frac{\partial \mu_m}{\partial z_k}]^* \quad m \geq 0 \quad (3.38)$$

$$\frac{\partial \mu_m}{\partial t} = [F_m(t, z)]_{z_1=0} \quad (3.39)$$

where the starred bracket means that z_1 must be replaced by $\frac{z_1}{\kappa_m}$ inside the bracket such that

$$\kappa_0 = \frac{z_1}{z_1 - \mu_0} \quad (3.40)$$

$$\kappa_1 = e^{\mu_1} \quad (3.41)$$

$$\kappa_{m+1} = \sqrt[m]{1 + m\mu_{m+1}^m z_1^m} \quad m \geq 1 \quad (3.42)$$

Let us now, seek a bound for $f_1^{(0)}, \dots, f_N^{(0)}$ by using the Cauchy-Schwartz inequality in the Eq(3.35)

$$\sum_{k=1}^N \sum_{l=1}^N b_{jkl} z_k z_l < \beta_j r^2 \quad (3.43)$$

where β_j represents the following sum

$$\beta_j = \sqrt{\sum_{k=1}^N \sum_{l=1}^N |b_{jkl}|^2} \quad (3.44)$$

Hence, we conclude

$$|f_j^{(0)}(t, \mathbf{z})| < \beta r^2 e^{-\lambda t} \quad (3.45)$$

where

$$\beta = \sqrt{\sum_{m=1}^N \beta_m^2} \quad (3.46)$$

Equation (3.45) implies that

$$|f_j^{(0)}(t, \mathbf{z} - \mu_0 \mathbf{e}_1)|_{z_1=0} < \beta \{|\mu_0|^2 + R^2\} e^{-\lambda t} \quad (3.47)$$

where

$$R = \sqrt{r^2 - |z_1|^2} \quad (3.48)$$

If we recall that μ_0 is a function of time and $(N - 1)$ space coordinates, z_2, \dots, z_N , then we can write the following inequalities via Lemma 1 and Eq(3.34)

$$|\mu_0(t; z_2, \dots, z_N)| < M_0(t, R) \quad (3.49)$$

$$\left| \frac{\partial \mu_0}{\partial t} \right| < \beta \{ \mu_0^2 + R^2 \} e^{-\lambda t} \left\{ 1 + (N - 1) \frac{\partial M_0}{\partial R} \right\} \quad (3.50)$$

We can obviously write that

$$M_0^2 + R^2 < (M_0 + R)^2 \quad (3.51)$$

and

$$\left| \frac{e^{-\lambda t}}{\beta} \frac{\partial \mu_0}{\partial t} \right| < \frac{\partial M_0}{\partial u} \quad (3.52)$$

where

$$u = \frac{\beta}{\lambda} [1 - e^{-\lambda t}] \quad (3.53)$$

By using all of these inequalities we can arrive at the following partial differential equation to produce the majorant function M_0 for μ_0

$$\frac{\partial(M_0 + R)}{\partial u} = (M_0 + R)^2 [1 + (\bar{N} - 1) \frac{\partial(M_0 + R)}{\partial R}] \quad (3.54)$$

The accompanying initial conditions for this nonlinear partial differential equation can be given in the following parametric form

$$M_0 = 0 \quad R = s_1 \quad u = 0 \quad (3.55a, b, c)$$

Although it is a nonlinear partial differential equation, its solution can be obtained via the method of characteristics. The equations for the characteristics are

$$\frac{\partial u}{\partial s_2} = 1 \quad (3.56a)$$

$$\frac{\partial R}{\partial s_2} = -(N - 1)(M_0 + R)^2 \quad (3.56b)$$

$$\frac{\partial(M_0 + R)}{\partial s_2} = (M_0 + R)^2 \quad (3.56c)$$

The solution of these equations together with Eqs(3.55a,b,c) give three parametric expressions $u = u(s_1, s_2)$; $R = R(s_1, s_2)$; $M_0 = M_0(s_1, s_2)$. The elimination of s_1 and s_2 among these three entities gives the following explicit expression

$$M_0 = \frac{1 - uR}{2(N - 1)u} \left[1 - \sqrt{1 - \frac{4(N - 1)uR}{(1 - u)^2}} \right] - R \quad (3.57)$$

Let us now, assume that we have constructed the following majorant functions for the f and F functions

$$|f_k^{(m)}| < \frac{\phi_k^{(m)}}{1 - \frac{\varepsilon_1}{\rho_m}} \quad k = 1, \dots, N \quad (3.58)$$

$$|F_m| < \frac{\Phi_m}{1 - \frac{\varepsilon_1}{\rho_m}} \quad (3.59)$$

where $\phi_k^{(m)}$, Φ_m and ρ_m depends on time and z_2, \dots, z_N . If we consider Eq(3.42), then we can produce the following manipulations

$$|f_k^{(m+1)}| < \frac{\phi_k^{(m)}}{1 - \frac{z_1}{\rho_m \kappa_m}} = \frac{\phi_k^{(m)} G_m}{1 - {}^{m-1}\sqrt{(m-1)|\mu_m| + \rho_m^{1-m} z_1}} \quad k \geq 2 \quad (3.60a)$$

$$G_m = \frac{1 - {}^{m-1}\sqrt{(m-1)|\mu_m| + \rho_m^{1-m} z_1}}{1 - \frac{z_1}{\rho_m \kappa_m}} \quad (3.60b)$$

As can be shown by a careful analysis, G_m is a decreasing function of z_1 and is bounded by unity as long as z_1^{m-1} remains smaller than $(m-1)|\mu_m| + \rho^{1-m}$. Hence

$$|f_k^{(m+1)}| < \frac{\phi_k^{(m+1)}}{1 - \frac{z_1}{\rho_{m+1}}} \quad (3.61)$$

which implies that

$$\phi_k^{(m+1)} = \phi_k^{(m)} \quad k \geq 2 \quad (3.62)$$

and

$$\rho_{m+1} = \frac{\rho_m}{{}^{m-1}\sqrt{1 + (m-1)|\mu_m| \rho_m^{m-1}}} \quad (3.63)$$

To obtain the recursion among the $\phi_1^{(m)}$ -parameters we need a little further analysis as follows

$$\left| \frac{f_k^{(m)} - \{f_k^{(m)}\}_{z_1=0}}{z_1} \right|^* < \frac{\phi_k^{(m)}}{\rho_m [1 - \frac{z_1}{\rho_m \kappa_m}]} < \frac{\phi_k^{(m)} G_m}{\rho_m [1 - \frac{z_1}{\rho_{m+1}}]} < \frac{\phi_k^{(m)}}{\rho_m [1 - \frac{z_1}{\rho_{m+1}}]} \quad k = 1, \dots, N \quad (3.64)$$

$$|f_1^{(m+1)}| < \frac{1}{1 - \frac{z_1}{\rho_{m+1}}} \left\{ \frac{\phi_1^{(m)}}{\rho_m} + \sum_{k=2}^N \frac{\phi_k^{(m)}}{\rho_m} \left| \frac{\partial \mu_m}{\partial z_k} \right| \right\} \quad (3.65)$$

If we use Lemma 1 and the following inequalities

$$\phi_k^{(m)} = \phi_k^{(0)} < \beta(R^2 + |\mu_0|^2) e^{-\lambda t} < \beta(R + |\mu_0|)^2 e^{-\lambda t}, \quad (3.66)$$

then we can conclude that

$$\phi_1^{(m+1)} = \frac{\phi_1^{(m)}}{\rho_m} + \frac{(N-1)\beta(R + M_0)^2 e^{-\lambda t}}{\rho_m} \frac{\partial M_m}{\partial R} \quad (3.67)$$

where $M_m(t, R)$ stands for the majorant function of μ_m . Therefore we have proved the following lemma:

LEMMA 2:

The $f_k^{(m)}$ -functions appearing in the construction of the μ -coefficients, are majorized by the following functions

$$|f_k^{(m)}| < \frac{\phi_m^{(m)}}{1 - \frac{z_1}{\rho_m}} \quad k = 1, \dots, N \quad (3.68)$$

where $\phi_m^{(k)}$ and ρ_m are functions of z_1, \dots, z_N and t , the definitions of which are given through the recursions presented in the Eqs(3.62), (3.63) and (3.67).

Let us consider again the recursion for ρ_n . If we assume that

$$\sqrt[n]{n\mu_{n+1}} e^{\lambda t} < \omega \quad (3.69)$$

then we can write

$$\alpha_n \leq e^{\lambda t} \rho_n \quad (3.70)$$

$$\alpha_{n+1} = \frac{1}{\sqrt[n+1]{1 + \omega \alpha_n^n}} \quad \alpha_1 = \rho_f \quad (3.71a, b)$$

where ρ_f can be determined from the quadratic structure of the descriptive functions. As can be shown after appropriate intermediate steps, α_n converges to a non-zero limit, say α , as n tends to infinity. This implies that

$$\lim_{n \rightarrow \infty} \rho_n = \rho \geq e^{\lambda t} \alpha > 0 \quad (3.72)$$

Since the sequence ρ_1, ρ_2, \dots , is a decreasing one we can change the recursion for $\phi_1^{(m)}$ as follows

$$\phi_1^{(m+1)} = \frac{\phi_1^{(m)}}{\rho} + \frac{(N-1)\beta(R + M_0)^2 e^{-\lambda t}}{\rho} \frac{\partial M_m}{\partial R} \quad (3.73)$$

This does not cause any remarkable difference in the construction of majorants except a possible decrease in the convergence radii of the majorant series. The explicit expression for the solution of the last difference equation can be expressed as follows

$$\phi_1^{(m)} = \frac{\phi_1^{(1)}}{\rho^{m-1}} + (N-1)\beta(R + M_0)^2 e^{-\lambda t} \left\{ \sum_{j=0}^{m-1} \frac{\partial M_j}{\partial R} \rho^{j-m} - \frac{\partial M_0}{\partial R} \rho^{-m} \right\} \quad (3.74)$$

Now, we can write the following equation through the above development

$$\left| \frac{\partial \mu_m}{\partial t} \right| < \frac{\phi_1^{(m)}}{\rho} + \frac{(N-1)\beta(R+M_0)^2 e^{-\lambda t}}{\rho} \frac{\partial M_m}{\partial R} < \frac{\phi_1^{(1)}}{\rho^m} + (N-1)\beta(R+M_0)^2 e^{-\lambda t} \left\{ \sum_{j=0}^m \frac{\partial M_j}{\partial R} \rho^{j-m} - \frac{\partial M_0}{\partial R} \rho^{-m} \right\} \quad (3.75)$$

By using the previously defined u-variable we can write

$$\frac{\partial M_m}{\partial u} = \frac{[1 - \frac{\lambda}{\beta} u] \phi_1^{(1)}}{\beta \rho^m} - (N-1) \frac{(R+M_0)^2}{\rho^{m+1}} \frac{\partial M_0}{\partial R} + (N-1)(R+M_0)^2 \sum_{j=0}^m \frac{\partial M_j}{\partial R} \rho^{j-m-1} \quad m \geq 1 \quad (3.76)$$

If we multiply this equation by $|\xi_1|^m e^{-m\lambda t}$ and sum both sides over m from 1 to ∞ , we obtain

$$\frac{\partial Z}{\partial u} = \left\{ \frac{[1 - \frac{\lambda}{\beta} u] \phi_1^{(1)}}{\beta} + (N-1) \frac{(R+M_0)^2}{\rho} \frac{\partial Z}{\partial R} \right\} \cdot \frac{1}{1 - \frac{|\xi_1|}{\rho} [1 - \frac{\lambda u}{\beta}]} \quad (3.77)$$

where

$$Z = \sum_{m=1}^{\infty} |\xi_1|^m [1 - \frac{\lambda u}{\beta}]^m M_m \quad (3.78)$$

Eq(3.77) is a first order linear partial differential equation for Z . If we consider the accompanying characteristics of this equation as

$$Z = 0, \quad u = 0, \quad R = t, \quad (3.79a, b, c)$$

then we can solve it via standard techniques and it is not difficult to show that the solution converges in a non-empty region of $\{u, R\}$ -space. Moreover, we can discard all the cases where a non-zero R exists, since we are able to bring all factorization problems into a canonical form. Hence we can replace R with 0 in our all the previous analysis and this yields

$$[\mu_m]_{R=0} = \sigma_m(t) \quad (3.80)$$

$$\phi_1^{(m+1)} = \frac{\phi_1^{(m)}}{\rho}; \quad \phi_1^{(m+1)} = \phi_1^{(m)} \frac{e^{-\lambda t}}{\alpha}; \quad \phi_1^{(m)} = \phi_1^{(1)} \frac{e^{-(m-1)\lambda t}}{\alpha^{m-1}} \quad (3.81)$$

$$|\sigma_{n+1}(t)| < \frac{\phi_1^{(1)}}{\alpha^n} \frac{1 - e^{-n\lambda t}}{n\lambda} \quad (3.82)$$

This last inequality provides the boundedness condition of $[n|\sigma_{n+1}|]^{1/n} e^{\lambda t}$ globally for $\lambda < 0$ and temporarily (conditionally) for $\lambda \geq 0$. These results can be summarized in the following theorem.

THEOREM 2:

If we consider a multidimensional system with quadratic descriptive functions which vanish at the origin and denote its characteristic mode by λ , then the following statements are true:

- i) If $\lambda < 0$, then the system is "Global Normal".
- ii) If $\lambda > 0$, then the system is at least "Temporary Normal".

Our third theorem is exactly the same of the one-dimensional case[2,3], and we give it without proof.

THEOREM 3:

If we define

$$\omega = \min_{1 \leq n \leq \infty} |n\sigma n + 1|^{-1/n} \quad (3.83)$$

and $|\xi_n|$ remains smaller than ω for a finite fixed n value, say N , then all higher order ξ -approximants also remain smaller than ω in absolute value.

An explicit expression of this theorem is as follows: "If the system under consideration is globally normal then the limit of the ξ -approximant sequence, $\xi(s, t) = \lim_{n \rightarrow \infty} \xi_n$, remains permanently in the main clean region of the multidimensional complex Euclidean space as time evolves".

These theorems imply that the best circumstance for the convergence of ξ - approximants is the case where λ is negative and ξ_1 is in the clean region. Since we have considerable flexibility permitted by the space extension transformations, we may affect λ by changing the convergence control parameters ν_1, \dots, ν_N , in such a way that our system becomes a globally normal system in a higher dimensional space. So we have the power to handle all factorization problems of augmented Lie evolution operators. Establishing this capability, we open the way for the development of associated software.

4. CONCLUDING REMARKS

In the first one of these two papers we showed that the most systems encountered in practical applications, can be brought into a quadratic canonical form via an appropriate space extension. We also constructed an extended transformation which made it possible to deal with canonical factorizations and permitted certain flexibilities, to affect the convergence properties of the resulting ξ -approximant sequences.

The first paper included the general formulation and the standardization of the scheme, and this paper presented the theorems about the convergence properties of the ξ -approximants. The most important result obtained here is the convergence properties of the factorization scheme. In other words, we may convert the system under consideration into the another one which has a characteristic mode with a negative real part. This opens up the possibility of dealing with global normal systems, however, the convergence control parameters, $\nu_1, \nu_2, \dots, \nu_N$, and the magnitude of λ determines the convergence radius. If the point where $\xi_1 = 1$ is outside of the main clean region then convergence failure may happen. However, our proofs are obtained under tight restrictions due to the utilization of the majorant series method. Hence the ξ -approximants may, very possibly, converge unless one of them encounters singularities of the mapping between that one and its higher order neighbour, due to the contractive mapping type of behavior of the recursion between them. Therefore, the convergence investigation for a given ξ -sequence on the entire complex plane will be an important step to take.

Finally we draw attention to the following cautionary comment. The possibility of changing one of the characteristic modes of system does not imply the possibility of changing of its asymptotic character when t tends to infinity. In other words, we can reveal the "Global Normality" of the system only when it really does exist. If the system under consideration has a composite structure such as only one part of its characteristic modes has negative real parts, then certain evolutions of the system can not have a "Global Normal" behaviour. In these case, the breakdown of the convergence or a convergence deceleration may be expected. Hence pre-knowledge about the characteristic modes seems to be quite useful.

ACKNOWLEDGEMENT

The authors would like to thank Professor Hilmi Demiray for helpful comments.

REFERENCES

- [1] M. Demiralp, H. Rabitz, 'Lie algebraic factorization of the multivariable evolution operators: Definition and the solution of the canonical problem" (*to be published*)
- [2] M. Demiralp, H. Rabitz, 'Factorization of certain evolution operators using Lie algebra: Formulation of the method" (*to be published*)
- [3] M. Demiralp, H. Rabitz, 'Factorization of certain evolution operators using Lie algebra: Convergence theorems (*to be published*)

Appendix L

12. Lie Algebraic Factorization of Multivariable Evolution Operators:
Definition and Solution of the Canonical Problem, M. Demiralp and H.
Rabitz, Int. J. of Eng. Sci., in press.

**LIE ALGEBRAIC FACTORIZATION OF MULTIVARIABLE
EVOLUTION OPERATORS: DEFINITION AND THE SOLUTION
OF THE CANONICAL PROBLEM***

Metin Demiralp and Herschel Rabitz**

Princeton University, Department of Chemistry
Princeton, N.J. 08544-1009, USA

* Supported by NATO via RG.86/0123, the Air Force Office of Scientific Research and the Office of Naval Research

** Permanent Address: İstanbul Technical University, Faculty of Sciences and Letters, Engineering Sciences Department, Ayazağa Campus, Maslak, 80626 - İstanbul, TURKEY

ABSTRACT

We have recently shown that the factorization of certain Lie algebraic evolution operators into a convergent infinite product of simple evolution operators is possible for one-dimensional cases. In this paper, we deal with the multivariable case. To this end, we formulate the factorization for the general case, then we show that the most of the practical problems can be brought to a canonical one. The canonical problem has nothing different in concept but the relevant partial differential equations to be solved can be easily handled. Two simple illustrative examples and the concluding remarks complete the work.

1. INTRODUCTION

All dynamical problems of physics and engineering can be characterized via properly defined evolution operators [1-4]. This is not only peculiar to classical mechanics; problems of quantum dynamics and non-equilibrium statistical mechanics [5-14] may also be treated through appropriate evolution operators. Most practically encountered problems necessitate the use of evolution operators in exponential form. Perhaps, the most important of these types is the Lie algebraic evolution operator which has a first order linear partial differential operator argument. There is, also, a close connection between the solution of first order differential equation systems as initial value problems and Lie algebraic evolution operators [4]. Hence, to establish a proper scheme to approximate the Lie algebraic evolution operators is of considerable importance. The resultant should be easily programmable such that it can be executed rapidly and require minimal memory. The efforts to approximate Lie algebraic evolution operators are not new. A well known early result is the Baker-Campbell-Hausdorff (BCH) formula where the product of two exponential operators is expressed in terms of various commutators between the arguments of these exponential operators [15-17], and the operators are not restricted to be Lie-algebraic ones.

In general, evolution operators have a tracing parameter which guides us when we develop a scheme to approximate them. Since time is the parameter which determines the point of the evolution, we can refer to this tracing parameter as time. However, we must keep in mind that certain exponential operators, like ones of the partition function in equilibrium statistical mechanics, may have same kind parameter but with a different physical meaning. A similar formula to BCH may be developed to approximate the exponential operators in an infinite product of exponentials such that each factor has a different integer power of tracing parameter in an increasing order [18]. In another context, operator techniques are often used to connect quantum mechanical entities with the classical ones [19-28]. Among these, Lie algebraic techniques have been investigated in most detailed manner [29-36]. The solution of the first order linear operator-differential equations with the aid of Lie algebraic methods or via commutator algebra has also been extensively studied. The use of the normal ordering of the operators provides a valuable means to solve these types of equations [8,10,11,37-42]. As mentioned above, exponential evolution

operators are also used in statistical mechanics. There, the arguments of the operators are different for classical and quantum mechanical cases, and generally, the main purpose is the evaluation of the partition function [43-46].

Powerful techniques are available to approximate the Lie algebraic exponential operators via Lie groups and via Lie algebraic theories [5-7,12,13]. These techniques are also used to calculate the classical mechanical trajectories of certain systems by using a prior known reference trajectory [1-3]. Since Lie algebra and Lie groups are frequently employed in mathematical physics, one can find many references to them in that literature [47-57].

As stated at the beginning of this section, the initial value problem of the first order differential equations system can be handled by using a vector field concept or Lie algebraic evolution operator. The evolution operators may be approximated as polynomial operators in terms of the argument of the evolution operator [4]. Although this approach gives quite accurate results in the initial period of the evolution, the discrepancy increases as time evolves due to unavoidable accumulations of errors.

With this information as background we desire an approximation scheme which globally characterizes the evolution under consideration. In other words, the scheme should be able to relate any point of the evolution to initial point without a knowledge about the other points. Hence, in earlier work we found a factorization scheme for Lie algebraic exponential evolution operators (LAEEOs) for one-variable cases [58,59]. As we have shown, LAEEO is expressed as an infinite product of simple evolution operators which can be handled easily and analytically. The action of the truncated products of this representation on a given function globally converges to a common limit which is the action of LAEEO on that given function. There are some restrictions on the convergence theorems given in those works. However these are sufficient conditions, so there still remains flexibility to extend the coverage of the theorems. This point will be investigated in our future works. Here, we generalize (and modify whenever it is necessary) the results of the one-variable case to multivariable cases.

The remainder of this paper is organized as follows. Section 2. gives the general formulation of the global factorization for multivariable systems followed by the explanation of the space extension concept and the definition of the canonical factorization problem in Section 3. The solution of the canonical factorization problem is given in Section 4.

A simple illustrative example and concluding remarks are presented in Section 5. and 6., respectively. The convergence properties of the scheme are given in the companion of this work.

2. FORMULATION OF THE FACTORIZATION SCHEME FOR THE MULTIVARIABLE CASES

A multivariable LAEEO can be written as follows

$$Q = e^{tL} \quad (2.1)$$

where L denotes a Lie operator defined as first order linear partial differential operator with

$$L = \sum_{j=1}^N f_j(z_1, z_2, \dots, z_N) \frac{\partial}{\partial z_j} \quad (2.2)$$

where f_j , $j = 1, 2, \dots, N$, are denoted as the descriptive functions of the system under consideration (e.g., the right hand side of a set of N coupled ordinary differential equations) and the number of variables is N . Although the variables are real in most practical cases, the z -variables are considered as complex valued to facilitate the proof of certain convergence theorems. The descriptive functions are assumed to be infinitely differentiable with respect to their arguments in the entire N -tuple complex space which is the cartesian product of the individual complex planes of the z -variables. Since many practical cases involve these types of descriptive functions, there is only a minor loss of generality. Indeed for most circumstances where the descriptive functions are infinitely differentiable for only certain subspaces of the N -tuple complex space of the z -variables, the problem can be altered via space extension transformations to satisfy the above assumption. We shall discuss the space extension concept later. A second assumption about the descriptive functions requires that they vanish when all the z -variables vanish. This assumption does not create any loss of generality since a space extension transformation can always assure this property to descriptive functions, as we shall see later.

We expand the descriptive functions to a multivariable Taylor series as follows

$$f_j(z_1, z_2, \dots, z_N) = \sum_{k=1}^{\infty} \sum_{l=1}^{n_k} a_{j,k}^{(l)} P_k^{(l)} \quad (2.3)$$

where $P_k^{(l)}$ stands for a multinomial* which belongs to the set of k -th degree homogeneous multinomials of the z -variables and its superscript, l characterizes its place in the set. The index n_k is the number of possible k -th degree homogeneous multinomials. The $a_{j,k}^{(l)}$ -coefficients define the system under consideration and are assumed to be known. In this text, we use the word "system" to characterize a point in the N -tuple complex space of the z -variables such that the motion of this point is completely specified when the descriptive functions are given. To define $P_k^{(l)}$ more explicitly we can write

$$P_k^{(l)} = z_1^{l_1} z_2^{l_2} \dots z_N^{l_N} \quad (2.4)$$

where $l_1 + l_2 + \dots + l_N = k$ and l is related to l_1, l_2, \dots, l_N through a function which takes integer values between 1 and n_k inclusive. The functional structure of this relation is completely arbitrary unless one specifies a scheme for the elements of k -th degree multinomials set. Utilizing Eq.(2.3) we can write the following expansion for our Lie operator

$$L = \sum_{j=1}^N \sum_{k=1}^{\infty} \sum_{l=1}^{n_k} a_{j,k}^{(l)} L_{j,k}^{(l)} \quad (2.5)$$

where

$$L_{j,k}^{(l)} = P_k^{(l)} \frac{\partial}{\partial z_j} \quad (2.6)$$

which may be called as "Fundamental Lie Operator". As can be easily shown, the commutator of any two fundamental Lie operators is again a fundamental Lie operator. In other words, the infinite set of fundamental Lie operators is closed under the commutation operation.

Now, we can construct fundamental evolution operators as below

$$Q_{j,k,l}^{(F)} = e^{\sigma(t) L_{j,k}^{(l)}} \quad (2.7)$$

where $\sigma(t)$ is assumed to be known. We call these operators "Fundamental Evolution Operators" because of the simplicity of the calculation of their action on a given infinitely differentiable function.

* We use the word "multinomial" instead of the word "polynomial" to imply multivariable polynomials

We now review certain fundamental properties of LAEEO's before attempting to find an explicit expression for the action of $Q_{j,k,l}^{(F)}$ on a given infinitely differentiable function of the z -variables. If g , h and Q_L denote two given functions of the z -variables and a general LAEEO respectively, we can write the following equations

$$Q_L\{gh\} = Q_L\{g\}Q_L\{h\} \quad (2.8)$$

$$Q_L g(z_1, z_2, \dots, z_N) = g(Q_L z_1, Q_L z_2, \dots, Q_L z_N) \quad (2.9)$$

where the first equation comes from the exponential structure of Q_L and the Leibnitz rule of the differentiation of a product. We call the second equation a "Penetration Property" and it can be derived via consecutive application of the first property on the multivariable Taylor expansion of g . We define Q_D as the simplest LAEEO which is called a "Displacement Operator" satisfying the following equation.

$$Q_D g(z_1, z_2, \dots, z_N) = \exp\left\{\sum_{j=1}^N \sigma_j \frac{\partial}{\partial z_j}\right\} g(z_1, z_2, \dots, z_N) = g(z_1 + \sigma_1, z_2 + \sigma_2, \dots, z_N + \sigma_N) \quad (2.10)$$

An examination of the structure of the fundamental evolution operators reveals that

$$Q_{j,k,l}^{(F)} z_m = 1 \quad m \neq j \quad (2.11)$$

Hence we can write

$$Q_{j,k,l}^{(F)} g(z_1, z_2, \dots, z_j, \dots, z_N) = g(z_1, z_2, \dots, z_{j-1}, Q_{j,k,l}^{(F)} z_j, z_{j+1}, \dots, z_N) \quad (2.12)$$

The last equation states that the action of a fundamental evolution operator on a given infinitely differentiable function of the z -variables is calculated through the action of the same operator on the z -variable which appears in the partial differential operator. To accomplish this task we can conveniently use the following entities

$$Z_j = z_j^{1-l_j} \quad \sigma^* = (1 - l_j) \sigma(t) z_1^{l_1} \dots z_{j-1}^{l_{j-1}} z_{j+1}^{l_{j+1}} \dots z_N^{l_N} \quad (2.13a, b)$$

and we simply obtain the following result

$$Q_{j,k,l}^{(F)} z_j = e^{\sigma^* \frac{\partial}{\partial z_j}} Z_j^{1/(1-l_j)} = (\sigma^* + Z_j)^{1/(1-l_j)} = z_j (1 + \sigma^* z_j^{l_j-1})^{1/(1-l_j)} \quad (2.14)$$

This equation remains valid for all non-negative integer values of l_j , however, the case where $l_j = 1$ necessitates a limiting procedure to obtain the following exponential structure

$$\{Q_{j,k,l}^{(F)}\}_{l_j} =: z_j = e^{\sigma'} z_j \quad \sigma' = \frac{\sigma^*}{(1 - l_j)} \quad (2.15a, b)$$

Now, we look at the meaning of the fundamental evolution operators. The first one of the Eqs.(2.14) implies that every fundamental evolution can be interpreted as a displacement transformation. The remaining Eqs.(2.14) have this interpretation. If we consider a set of functions within which every member is continuous and square integrable along a given finite path in the complex plane of Z_j and vanishes at the endpoints of the same path, then we can easily show that $Q_{j,k,l}^{(F)}$ is a self-adjoint operator on this set. Hence, every fundamental evolution operator corresponds to a rotation in a properly defined Hilbert space, so they may be considered as unitary transformations. The fundamental evolution operators play a role like $L_{j,k}^{(l)}$ does in the multivariable Taylor expansion when we attempt to factorize LAEEO. In other words, we can write the following factorization equation with proper choices of each individual σ -coefficient appearing in the fundamental evolution operators

$$Q = \prod_{j,k,l} Q_{j,k,l}^{(F)} \quad (2.16)$$

where we have not specified a particular ordering of the factors. However, all possible fundamental evolution operators are included in the product. The validity of above equation can be shown via closed property of the set of fundamental evolution operators under commutation operation. The ordering arbitrariness appearing in the factorization formula above gives us the opportunity of constructing the simplest factorization scheme. Since the action of Q on a given function necessitates only the calculation of the individual actions of Q on the z -variables, we can deal with the calculation of Qz_1 for simplicity. Indeed, we can obtain the value of Qz_j by simply interchanging the roles of z_1 and z_j . Hence, our main task is rather to evaluate the action of LAEEO on z_1 . Now to write a specific factorization formula, we can use the following criteria:

i) Factors which include differentiation with respect to same variable must be collected in the same group. This creates N different groups of factors.

ii) Every group of factors must be composed of subgroups such that every factor of a specific subgroup must have the factors which have the homogeneous multinomials of same degree.

iii) Each of these subgroups of factors can be considered as a single fundamental evolution operator.

iv) The factors corresponding to linear multinomials must be collected as a single leftmost factor. This is simply separation of the linear response of the system under consideration and is useful for certain computational purposes (for example, it may reduce the accumulation of errors). Therefore we can propose the following factorization formula

$$Qz_1 = Q_L \prod_{j=1}^N \left\{ \prod_{k=0}^{\infty} e^{\mu_k^{(j)} z_j^k \frac{\partial}{\partial z_j}} \right\} z_1 \quad (2.17)$$

where $\mu_k^{(j)}$ depends on t and all the z -variables except z_j , and Q_L is defined as

$$Q_L = \exp \left\{ t \sum_{l=1}^N \sum_{k=1}^N a_{l,1}^{(k)} z_k \frac{\partial}{\partial z_l} \right\} \quad (2.18)$$

The consecutive actions of the last $N - 1$ infinite products of Eq(2.17) on z_1 produce no change on it. Hence we can simply discard them and drop the superscript of the undetermined coefficient, $\mu_k^{(j)}$. Therefore the factorization takes on the form

$$Qz_1 = Q_L \left\{ \prod_{k=0}^{\infty} e^{\mu_k z_1^k \frac{\partial}{\partial z_1}} \right\} z_1 \quad (2.19)$$

Now, first we have to find a practical way to approximate Qz_1 and second, we have to relate the undetermined μ_k -coefficients to the descriptive functions of the system under consideration. The first item can be handled by defining the following " $\bar{\xi}$ -approximants" in analogy with earlier work [58,59]

$$\bar{\xi}_n = Q_L \left\{ \prod_{k=0}^n e^{\mu_k z_1^k \frac{\partial}{\partial z_1}} \right\} z_1 \quad n = 0, 1, \dots \quad (2.20)$$

The over bar will be dropped when we change the definition of these approximant into a more efficient form for computational purposes. These approximants can be recursively determined in the following way

$$\bar{\xi}_{n+1} = Q_L \left\{ \prod_{k=0}^n e^{\mu_k z_1^k \frac{\partial}{\partial z_1}} \right\} e^{\mu_{n+1} z_1^{n+1} \frac{\partial}{\partial z_1}} z_1 \quad n = 0, 1, \dots \quad (2.21a)$$

$$e^{\mu_{n+1} z_1^{n+1} \frac{\partial}{\partial z_1}} z_1 = z_1 (1 - n \mu_{n+1} z_1^n)^{-1/n} \quad n = 0, 1, \dots \quad (2.21b)$$

$$\bar{\xi}_{n+1} = \bar{\xi}_n (1 - n \mu_{n+1} \bar{\xi}_n^n)^{-1/n} \quad n = 0, 1, \dots \quad (2.22)$$

where we have used the penetration property of LAEEO's consecutively and the definition of the $\bar{\xi}$ -approximants. Although this recursive relation is first order, it is non-linear and has a quite singular behaviour. If we consider the infinite set of the complex planes of $\bar{\xi}_n$; $n = 0, 1, 2, \dots$ we can interpret the recursion relations above as mappings between two consecutive member of this set. Since every mapping has a different order of algebraic branch point which moves in its plane as time evolves, then the limiting plane, $\bar{\xi}_\infty$, has an infinite number of moving trajectories of every order of algebraic branch points. Hence, the value of Qz_1 which can be considered* as $\bar{\xi}_\infty$ may have a quite singular dynamical structure depending on the values of the z -variables and time. Since the location of the branch point trajectories are completely determined by μ_k -parameters we can call them "Generators". Now, we have to give an explicit expression for $\bar{\xi}_0$ to establish the uniqueness of the $\bar{\xi}$ -approximants. We can write the following equalities for this purpose.

$$A_{jk} = a_{k,1}^j, \quad Q_L = e^{t\mathbf{z}^T \mathbf{A}^T \nabla} \quad (2.23a, b)$$

where \mathbf{z}, ∇ stand for the position vector and the gradient operator in the space spanned by z -variables and \mathbf{A} denotes the matrix which elements are given above. A careful investigation immediately shows that

$$Q_L \mathbf{z} = e^{t\mathbf{A}} \mathbf{z} \quad (2.24)$$

Therefore

$$\bar{\xi}_0 = \mathbf{e}_1^T e^{t\mathbf{A}} \mathbf{z} \quad (2.25)$$

where \mathbf{e}_1 denotes the first cartesian unit vector $[1, 0, \dots, 0]$.

Our second task, the determination of μ -functions, necessitates more detailed analysis. To this end, we can use the following superoperator equation for Q

$$\frac{\partial Q}{\partial t} = \sum_{j=1}^N f_j(\mathbf{z}) \frac{\partial Q}{\partial z_j} \quad [Q]_{t=0} = I \quad (2.26a, b)$$

* We shall prove this in the companion of this paper.

and we can draw on the linear response property as follows

$$Q = Q_L Q^{(0)}, \quad \frac{\partial Q^{(0)}}{\partial t} = \{-z^T A^T \nabla + \sum_{j=1}^N f_j(e^{-tA^T} z) Q_L^{-1} \frac{\partial}{\partial z_j} Q_L\} Q^{(0)},$$

$$[Q^{(0)}]_{t=0} = I \quad (2.27a, b, c)$$

where we have imposed the initial condition for $Q^{(0)}$ to preserve its unitarity (I stands for identity operator) at the beginning of the evolution and have used the penetration property of LAEEO's. There are unusually complicated operators in the right hand side of the Eq(2.27b). We simplify them by using the following identity based on the commutativities of the involved operators

$$Q_L^{-1} z^T A^T \nabla Q_L = z^T A^T \nabla \quad (2.28)$$

and, same identity can be written as follows via certain properties of LAEEO's

$$z^T e^{-tA^T} A^T Q_L^{-1} \nabla Q_L = z^T A^T \nabla \quad (2.29)$$

which implies

$$Q_L^{-1} \nabla Q_L = e^{tA^T} \nabla \quad (2.30)$$

Therefore we can write

$$\frac{\partial Q^{(0)}}{\partial t} = \left\{ \sum_{j=1}^N f_j^{(0)}(z, t) \frac{\partial}{\partial z_j} \right\} Q^{(0)} \quad [Q^{(0)}]_{t=0} = I \quad (2.31a, b)$$

where

$$f_j^{(0)}(z, t) = \sum_{k=1}^N Q_j^{(A)} f_j(Q^{(A)} z) \quad j = 1, \dots, N \quad Q^{(A)} = e^{-tA} \quad (2.32a, b)$$

Now, we can similarly extract the effect of the first factor of the infinite product in Eq(2.19) as follows

$$Q^{(0)} = e^{\mu_0 \frac{\partial}{\partial z_1}} Q^{(1)}, \quad \frac{\partial Q^{(1)}}{\partial t} = \left\{ \sum_{j=1}^N f_j^{(0)}(z_1 - \mu_0, z_2, z_3, \dots, z_N, t) e^{-\mu_0 \frac{\partial}{\partial z_1}} \frac{\partial}{\partial z_j} e^{\mu_0 \frac{\partial}{\partial z_1}} \right\} Q^{(1)}$$

$$(2.33a, b)$$

$$[Q^{(1)}]_{t=0} = I \quad (2.33c)$$

We can trace the following steps to simplify this equation.

$$\mu_0 \frac{\partial}{\partial z_1} \frac{\partial}{\partial z_k} - \frac{\partial}{\partial z_k} \mu_0 \frac{\partial}{\partial z_1} = - \frac{\partial \mu_0}{\partial z_k} \frac{\partial}{\partial z_1} \quad (2.34a)$$

\Rightarrow

$$(-\mu_0 \frac{\partial}{\partial z_1})^n \frac{\partial}{\partial z_k} - \frac{\partial}{\partial z_k} (-\mu_0 \frac{\partial}{\partial z_1})^n = \frac{\partial \mu_0}{\partial z_k} \frac{\partial}{\partial z_1} (-\mu_0 \frac{\partial}{\partial z_1})^{n-1} \quad (2.34b)$$

\Rightarrow

$$e^{-\mu_0 \frac{\partial}{\partial z_1}} \frac{\partial}{\partial z_k} e^{\mu_0 \frac{\partial}{\partial z_1}} = \frac{\partial}{\partial z_k} + \frac{\partial \mu_0}{\partial z_k} \frac{\partial}{\partial z_1} \quad k = 1, 2, \dots, N \quad (2.35)$$

\Rightarrow

$$\frac{\partial Q^{(1)}}{\partial t} = \{f_1^{(1)}(z, t) z_1 \frac{\partial}{\partial z_1} + \sum_{k=1}^N f_k^{(1)}(z, t) \frac{\partial}{\partial z_k}\} Q^{(1)} \quad [Q^{(1)}]_{t=0} = I \quad (2.36a, b)$$

where

$$z_1 f_1^{(1)}(z_1, z_2, \dots, z_N, t) = f_1^{(0)}(z_1 - \mu_0, z_2, z_3, \dots, z_N, t) + \sum_{k=2}^N f_k^{(0)}(z_1 - \mu_0, z_2, z_3, \dots, z_N, t) \frac{\partial \mu_0}{\partial z_k} - \frac{\partial \mu_0}{\partial t} \quad (2.37a)$$

$$f_k^{(1)}(z_1, z_2, \dots, z_N, t) = f_k^{(0)}(z_1 - \mu_0, z_2, z_3, \dots, z_N, t) \quad k = 2, 3, \dots, N \quad (2.37b)$$

Since we have extracted the factor which includes $\mu_0 \frac{\partial}{\partial z_1}$ from the infinite product representation of Q , $Q^{(1)}$ involves the remaining operators which vanish when z_1 goes to zero. Hence $f_1^{(1)}$ must be finite when $z_1 = 0$. This, however, implies that the right hand side of Eq(2.37a) must vanish when z_1 tends to zero and gives the following partial differential equation for μ_0

$$\frac{\partial \mu_0}{\partial t} = f_1^{(0)}(-\mu_0, z_2, z_3, \dots, z_N, t) + \sum_{k=2}^N f_k^{(0)}(-\mu_0, z_2, z_3, \dots, z_N, t) \frac{\partial \mu_0}{\partial z_k} \quad (2.38)$$

Now, if we define

$$F_0(z_1, \dots, z_N, t) = f_1^{(0)}(z_1 - \mu_0, z_2, \dots, z_N, t) + \sum_{k=2}^N f_k^{(0)}(z_1 - \mu_0, z_2, \dots, z_N, t) \frac{\partial \mu_0}{\partial z_k} \quad (2.39)$$

we can write $f_1^{(1)}$ in a more compact form

$$f_1^{(1)}(z_1, \dots, z_N, t) = \frac{F_0(z_1, z_2, \dots, z_N, t) - F_0(0, z_2, \dots, z_N, t)}{z_1} \quad (2.40)$$

We can extract the remaining factors of the infinite product representation of Q in this fashion and write the following superoperator equation for $Q^{(n)}$ which does not involve the $z_1^k \frac{\partial}{\partial z_1}$ operators for $k = 0, 1, \dots, n-1$.

$$\frac{\partial Q^{(n)}}{\partial t} = \{f_1^{(n)}(z, t) z_1^n \frac{\partial}{\partial z_1} + \sum_{k=1}^N f_k^{(n)}(z, t) \frac{\partial}{\partial z_k}\} Q^{(n)} \quad [Q^{(n)}]_{t=0} = I \quad (2.41a, b)$$

Then, we can proceed in the following way to obtain the superoperator equation for $Q^{(n+1)}$ by using certain properties of fundamental evolution operators.

$$Q^{(n)} = e^{\mu_n z_1^n \frac{\partial}{\partial z_1}} Q^{(n+1)} \quad (2.42)$$

$$\begin{aligned} \frac{\partial Q^{(n+1)}}{\partial t} &= \{ [f_1^{(n)}(\kappa_n z_1, z_2, \dots, z_N, t) - \frac{\partial \mu_n}{\partial t}] z_1^n \frac{\partial}{\partial z_1} + \\ &\sum_{k=2}^N f_k^{(n)}(\kappa_n z_1, z_2, \dots, z_N, t) e^{-\mu_n z_1^n \frac{\partial}{\partial z_1}} \frac{\partial}{\partial z_k} e^{\mu_n z_1^n \frac{\partial}{\partial z_1}} \} Q^{(n+1)} \\ [Q^{(n+1)}]_{t=0} &= I \end{aligned} \quad (2.43a, b, c)$$

where

$$\kappa_n = (1 + (n-1)\mu_n z_1^{n-1})^{-1/(n-1)} \quad (2.44)$$

and we have used some properties of the fundamental evolution operators. A careful analysis shows that

$$e^{-\mu_n z_1^n \frac{\partial}{\partial z_1}} \frac{\partial}{\partial z_k} e^{\mu_n z_1^n \frac{\partial}{\partial z_1}} = \frac{\partial}{\partial z_k} + \frac{\partial \mu_n}{\partial x_k} z_1^n \frac{\partial}{\partial z_1} \quad (2.45)$$

Therefore we can write

$$\begin{aligned} \frac{\partial Q^{(n+1)}}{\partial t} &= \{f_1^{(n+1)}(z, t) z_1^{n+1} \frac{\partial}{\partial z_1} + \sum_{k=2}^N f_k^{(n+1)}(z, t) \frac{\partial}{\partial z_k}\} Q^{(n+1)} \\ [Q^{(n+1)}]_{t=0} &= I \end{aligned} \quad (2.46a, b)$$

where

$$z_1 f_1^{(n+1)}(z_1, z_2, \dots, z_N, t) = f_1^{(n)}(\kappa_n z_1, z_2, z_3, \dots, z_N, t) + \sum_{k=2}^N f_k^{(0)}(\kappa_n z_1, z_2, z_3, \dots, z_N, t) \frac{\partial \mu_n}{\partial z_k} - \frac{\partial \mu_n}{\partial t} \quad (2.47a)$$

$$f_k^{(n+1)}(z_1, z_2, \dots, z_N, t) = f_k^{(n)}(\kappa_n z_1, z_2, z_3, \dots, z_N, t) \quad k = 2, 3, \dots, N \quad (2.47b)$$

However, the finiteness of $f_1^{(n+1)}$ for $z_1 = 0$ implies that

$$\frac{\partial \mu_n}{\partial t} = f_1^{(n)}(0, z_2, z_3, \dots, z_N, t) + \sum_{k=2}^N f_k^{(n)}(-\mu_0, z_2, z_3, \dots, z_N, t) \frac{\partial \mu_n}{\partial z_k} \quad (2.48)$$

Now, if we define

$$F_n(z_1, \dots, z_N, t) = f_1^{(n)}(\kappa_n z_1, z_2, \dots, z_N, t) + \sum_{k=2}^N f_k^{(n)}(\kappa_n z_1, z_2, \dots, z_N, t) \frac{\partial \mu_n}{\partial z_k} \quad (2.49)$$

we can write $f_1^{(n+1)}$ in a more compact form

$$f_1^{(n+1)}(z_1, \dots, z_N, t) = \frac{F_n(z_1, z_2, \dots, z_N, t) - F_n(0, z_2, \dots, z_N, t)}{z_1} \quad (2.50)$$

Therefore, we can compactly express all the discussions of this section in the following theorem.

THEOREM 1:

If we consider an N -variable system with given descriptive functions, $\{f_j(z_1, z_2, \dots, z_N) \mid j = 1, 2, \dots, N\}$ and consider its LAEEO,

$$Q = \exp\left\{t \sum_{j=1}^N f_j(\mathbf{z}) \frac{\partial}{\partial z_j}\right\} \quad (2.51)$$

then we can write the following factorization formula

$$Q z_1 = Q_L \left\{ \prod_{k=0}^{\infty} e^{\mu_k z_1 \frac{\partial}{\partial z_1}} \right\} z_1 \quad (2.52)$$

where

$$Q_L = \exp\left\{t \sum_{l=1}^N \sum_{k=1}^N a_{l,1}^{(k)} z_k \frac{\partial}{\partial z_l}\right\} \quad (2.53)$$

iff the following partial differential equation is satisfied by the μ -parameters

$$\frac{\partial \mu_n}{\partial t} = f_1^{(n)}(0, z_2, z_3, \dots, z_N, t) + \sum_{k=2}^N f_k^{(n)}(-\mu_0, z_2, z_3, \dots, z_N, t) \frac{\partial \mu_n}{\partial z_k}$$

$$\mu_n(z_2, \dots, z_N, 0) = 0 \quad n \geq 0 \quad (2.54a, b)$$

where

$$f_1^{(n+1)}(z_1, \dots, z_N, t) = \frac{F_n(z_1, z_2, \dots, z_N, t) - F_n(0, z_2, \dots, z_N, t)}{z_1} \quad n \geq 0 \quad (2.55)$$

$$F_n(z_1, \dots, z_N, t) = f_1^{(n)}(\kappa_n z_1, z_2, \dots, z_N) + \sum_{k=2}^N f_k^{(n)}(\kappa_n z_1, z_2, \dots, z_N, t) \frac{\partial \mu_n}{\partial z_k} \quad n \geq 0 \quad (2.56)$$

$$\kappa_n = (1 + (n-1)\mu_n z_1^{n-1})^{-1/(n-1)} \quad n \geq 0 \quad (2.57)$$

$$f_j^{(0)}(z, t) = \sum_{k=1}^N Q_{j,k}^{(A)} f_j(Q^{(A)} z) \quad j = 1, \dots, N \quad Q^{(A)} = e^{-tA} \quad (2.58a, b)$$

Hence, the factorization problem mainly reduces to the solution of an infinite number of partial differential equations. This may seem to be a forbidden task, however by the use of the space extension concept the matter can be brought to a level where necessary information can be easily obtained without attempting the solution of the partial differential equations given above. Then we shall see that the factorization problem can be transformed to an easier one such that the equations for μ -parameters can be handled in a finite number procedures. We shall discuss this later in a detailed manner.

Theorem 1 gives a necessary condition for the existence of the factorization. The sufficient conditions can be found when we deal with the convergence properties of the scheme. A quite detailed investigation is given in the companion to this paper.

3. SPACE EXTENSION CONCEPT AND THE DEFINITION OF THE CANONICAL PROBLEM

In the previous section, we developed the main aspects of the factorization scheme. There are some difficulties which may prevent bringing the scheme into a truly practical level. These may be gathered in the following three groups:

i) First of all, the descriptive functions, $\{f_j\}$ seem to contain an infinite number of parameters since they are represented in a power series. This necessitates an infinite amount of input information for the algorithm and is important for computational reasons. In fact, the descriptive functions encountered in the practical cases have only a few independent input parameters even if they can only be represented in power series. However, even in this case, there may be slow convergence due to the fact that the terms at a given order in the power series affect the further factors of the infinite product of the scheme. If one deal with the descriptive functions in global manner, these types of problems can be handled more easily.

ii) The second difficulty concerns the structure of linear response matrix which is given by Eq(2.23a). Undesired complications may arise since a matrix will generally rotate the position vector in addition to the changing its magnitude (an extension or contraction). Hence, the most preferable matrix to appear in the linear response terms is the identity matrix, I .

iii) The third difficulty involves the position of the factorization point. The factorization point corresponds to the initial conditions of the differential equations system. In the factorization scheme we used it in an implicit manner. The factorization point can be explicitly revealed by writing the factorization formula as follows

$$Q_{z_1} = \{Q_L \{ \prod_{k=0}^{\infty} e^{\mu_k \zeta_1^k \frac{\partial}{\partial \zeta_1}} \} \zeta_1\}_{\zeta=z} \quad (3.1a)$$

$$Q_L = \exp \{ t \sum_{l=1}^N \sum_{k=1}^N a_{l,1}^{(k)} \zeta_k \frac{\partial}{\partial \zeta_l} \} \quad (3.1b)$$

where ζ and z are utilized to represent the collections of the N -members of the corresponding entities. In this formula, z characterizes the factorization point and ζ stands for dummy variables employed not to confuse the intermediate differentiations. The complications arising from the factorization point arise in the calculations of μ -parameters. We recall that the μ -parameters depend on time and on the variables z_2, z_3, \dots, z_N . If the factorization point were the point where all z -variables except z_1 vanish, then the partial differential equations for the determination of μ s would be quite easily handled. This can be done when the vector z is an eigenvector of linear response matrix A if we use a

rotation transformation via an orthonormal matrix. Thus we will have $A = \lambda I$ and there will be no longer a problem with the position of the factorization point.

The descriptive functions for systems of the practical importance are generally expressible as multinomials of certain known functions of z_1, z_2, \dots, z_N . In mathematical language

$$f_j(z) = \sum \alpha_{k_1, k_2, \dots, k_{m_j}}^{(j)} \phi_1^{k_1}(z) \phi_2^{k_2}(z) \dots \phi_{m_j}^{k_{m_j}}(z) \quad j = \leq N \quad (3.2)$$

where the summation is carried over the k -values which satisfy $k_1 + k_2 + \dots + k_{m_j} \leq D_j$ (D_j is the degree of the multinomial for $f_j(z)$). The ϕ -functions above are assumed to be known functions such as polynomic, trigonometric, hyperbolic, logarithmic or hypergeometric and generalized hypergeometric functions. Any is appropriate for our purposes, but the choice of the ϕ -functions is not completely arbitrary. The set of ϕ -functions must have a finite number of elements and it must be closed under the action of the gradient operator with respect to the z -variables. If we denote the number of the members of this set by M then we can define the following new variables

$$w_k = \phi_k(z) \quad k = 1, \dots, M \quad (3.3)$$

and reexpress the Lie operator of the system under consideration as follows

$$L = \sum_{k=1}^M \left\{ \sum_{j=1}^N f_j(z) \frac{\partial \phi_k}{\partial z_j} \right\} \frac{\partial}{\partial w_k} \quad (3.4)$$

Since the terms inside the braces can be expressed as the multinomials of the w -variables, the problem reduces to the factorization of LAEEO of a system which has descriptive functions as multinomials in the system-variables. Therefore we change the phase space spanned by the z -coordinates to a new one spanned by the w -coordinates and the dimension of the space is also changed unless $M = N$. In most practical applications $M > N$, hence we call the change of space as "Space Extension Transformation". Although certain limited cases may have a lower dimensional space after the transformation* takes place, we shall use the word "Extension" with this comment in mind. Now, we can express the Lie operator of the system under consideration more explicitly as

$$\sum_{k=1}^N f_k(z) \frac{\partial \phi_j}{\partial z_j} = \sum_{k=1}^{D_s} \sum_{l=1}^{n_k} \beta_{j,k}^{(l)} P_k^{(l)} \quad j = 1, \dots, M \quad (3.5)$$

* which is namely a "Contraction"

where D_S is the maximum degree of the multinomials appearing in the new descriptive functions of the system on the extended space. The β -coefficients are given with the system and $P_k^{(l)}$ stands for the homogeneous multinomial in w -variables as follows

$$P_k^{(l)} = w_1^{l_1} w_2^{l_2} \dots w_M^{l_M} \quad (3.6)$$

and l -indices have the same meanings as in Eq(2.4). Therefore, the number of parameters to specify the descriptive functions is reduced to a finite value. However, there is still a possibility of further simplification in the structure of descriptive functions. Indeed, one can define the following new variables for this purpose

$$\omega_J = P_k^{(l)} \quad J = l + \sum_{j=1}^{k-1} n_j \quad 1 \leq J \leq \overline{M} = \sum_{j=1}^{D_S} n_j \quad (3.7)$$

Then, the descriptive functions become linearly dependent on the ω -variables. In addition, the action of the gradient operator with respect to the w -variables on any homogeneous multinomial represented by $P_k^{(l)}$ creates a linear combination of various homogeneous multinomials. Hence, the new descriptive functions of the system under consideration will be quadratic functions of the ω -variables and this is the smallest degree which can be taken by descriptive functions unless they are linear in the w -variables. All these matters are compactly given in the following Lemma.

LEMMA 1:

If the descriptive functions of a given system can be multinomially expressed in terms of the members of a finite set of functions which is closed under the action of the gradient operator, then one can find an appropriate space extension transformation which converts the system to another one which has quadratic descriptive functions in the new space coordinates.

Therefore, we can assume that the descriptive functions of a system can be expressed as follows

$$f_j = \alpha_j^{(0)} + \sum_{k=1}^N \alpha_{jk}^{(1)} z_k + \sum_{k=1}^N \sum_{l=1}^N \alpha_{jkl}^{(2)} z_k z_l \quad j = 1, \dots, n \quad (3.8)$$

where the α -constants are assumed to be given with the system and we return to use our original symbols for simplicity.

The quadratic structure of Eq.(3.8) is quite simple. However, a constant term arises which contrasts with the fundamental assumptions of the factorization of LAEEO's, hence we seek a new transformation which removes the constant terms in the structure of descriptive functions. It will be a significant simplification if the same transformation makes it possible to replace the linear response matrix $\alpha_{jk}^{(1)}$ with the identity matrix \mathbf{I} . Fortunately it is possible to find such transformations. To this end, we can define the following new variables

$$W_j = z_j \quad j = 1, \dots, N \quad (3.9a)$$

$$W_{N+1} = 1 \quad (3.9b)$$

Since $W_{N+1} = 1$ at the factorization point of the space spanned by the W -coordinates, we can simply multiply each of the constant terms in Eq(3.8) with W_{N+1} . This gives

$$f_j(\mathbf{W}) = \alpha_j^{(0)} W_{N+1} + \sum_{k=1}^N \alpha_{jk}^{(1)} W_k + \sum_{k=1}^N \sum_{l=1}^N \alpha_{jkl}^{(2)} W_k W_l \quad j = 1, \dots, N \quad (3.10a)$$

$$f_{N+1}(\mathbf{W}) = 0 \quad (3.10b)$$

which has no more constant terms and fulfills the fundamental assumption of the factorization scheme. Following the same reasoning, the vanishing property of the factor $W_{N+1} - 1$ at the factorization point, enables us to replace the Lie operator with the following one

$$L_{EX} = \sum_{j=1}^{N+1} f_j(\mathbf{W}) \frac{\partial}{\partial W_j} + L_R \quad (3.11a)$$

$$L_R = (W_{N+1} - 1) \sum_{j=1}^{N+1} \sum_{k=1}^{N+1} \gamma_{jk} W_k \frac{\partial}{\partial W_j} \quad (3.11b)$$

Indeed, if we properly use the commutator algebra between L and L_R we can show that the all of the terms resulting various commutation operations between L and L_R have a factor as $W_{N+1} - 1$. Hence we can easily prove the following lemma.

LEMMA 2:

The Lie operator of every quadratic system can be replaced with an extended one given in Eqs(3.11a,b)

As we mentioned above the residual operator, L_R has no contribution to the evolution on the factorization point where $W_{N+1} = 1$. However it permits us to change linear response matrix into the form we desire. It is sufficient to give the following specific values to γ_{jk}

$$\gamma_{jk} = (1 - \delta_{j, N+1})\{(\lambda\delta_{jk} - \alpha_{jk}^{(1)})(1 - \delta_{k, N+1}) - \alpha_j^{(0)}\delta_{k, N+1}\} \quad j, k = 1, \dots, N+1 \quad (3.12)$$

where δ_{jk} represents the usual Kroenecker's symbol, and λ stands for an undetermined parameter which may aid in adjusting the numerical convergence rate of the scheme. Therefore, the descriptive functions take the following forms

$$f_j(\mathbf{W}) = \lambda W_j + \sum_{k=1}^{N+1} \sum_{l=1}^{N+1} \bar{b}_{jkl} W_k W_l$$

$$j = 1, \dots, N+1 \quad (3.13)$$

where the \bar{b}_{jkl} -coefficients take the following values

$$\bar{b}_{jkl} = (1 - \delta_{j, N+1})(1 - \delta_{k, N+1})(1 - \delta_{l, N+1})\alpha_{jkl}^{(2)} + \gamma_{jk}\delta_{l, N+1}$$

$$j, k, l = 1, \dots, N+1 \quad (3.14)$$

The present Lie operator of the system and the factorization point can be written as follows

$$L = \sum_{j=1}^{N+1} \lambda W_j \frac{\partial}{\partial W_j} + \sum_{j=1}^{N+1} \sum_{k=1}^{N+1} \sum_{l=1}^{N+1} \bar{b}_{jkl} W_k W_l \frac{\partial}{\partial W_j} \quad (3.15)$$

$$W_j^{(f)} = z_j(1 - \delta_{j, N+1}) + \delta_{j, N+1} \quad j = 1, \dots, N+1 \quad (3.16)$$

where the z -symbols are no longer variable; they are just fixed values which represent a given specific point in the space of the W -coordinates together with the unit value of W_{N+1} . The last thing to do is the standardization of the factorization point. To this end, we consider an orthonormal set of $(N+1)$ -dimensional unit vectors $\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \dots, \mathbf{W}^{(N+1)}$ such that $\mathbf{W}^{(1)}$ is proportional to $\mathbf{W}^{(f)}$ and define a transformation matrix \mathbf{T} as below

$$\mathbf{T} = \overline{\mathbf{W}} \overline{\mathbf{T}} \quad (3.17)$$

$$\bar{\mathbf{W}} = [\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \dots, \mathbf{W}_{(N+1)}] \quad (3.18)$$

The non-zero elements of $\bar{\mathbf{T}}$ are given below

$$\bar{T}_{11} = s \quad \bar{T}_{kk} = 1 \quad \bar{T}_{1k} = s\nu_{k-1} \quad 2 \leq k \leq N+1 \quad (3.19a, b, c)$$

$$s = \{1 + \sum_{j=1}^{N+1} |z|_j^2\}^{1/2} \quad (3.19d)$$

where the ν -coefficients are certain complex values which enable us to evaluate the action of LAEEO on any linear combination of the coordinates as we shall see later. Then, we can transform our variables as follows

$$W_j = \sum_{k=1}^{N+1} T_{jk} \eta_k \quad j = 1, \dots, N+1 \quad (3.20)$$

and obtain the following Lie operator in the η -variables

$$L = \sum_{j=1}^{N+1} \lambda \eta_j \frac{\partial}{\partial \eta_j} + \sum_{j=1}^{N+1} \sum_{k=1}^{N+1} \sum_{l=1}^{N+1} b_{jkl} \eta_j \eta_k \frac{\partial}{\partial \eta_l} \quad (3.21)$$

where b_{jkl} satisfies

$$\sum_{m=1}^{N+1} T_{jm} b_{mkl} = \sum_{m=1}^{N+1} \sum_{n=1}^{N+1} \bar{b}_{jmn} T_{mk} T_{nl} \quad j, k, l = 1, \dots, N+1 \quad (3.22)$$

and the factorization point is, now, given as below

$$\eta_1^{(f)} = 1 \quad \eta_k^{(f)} = 0 \quad k = 2, 3, \dots, N+1 \quad (3.23)$$

This form of the our factorization problem is the simplest one unless the b_{jkl} -coefficients are specifically equal to zero. We refer this form as the "Canonical Factorization Problem of LAEEO's" or simply "Canonical Problem". Now, we can close this section by giving the following theorem which summarizes the discussions presented here.

THEOREM 2:

If the descriptive functions of a given system are multinomially expressible in terms of the members of a finite function set which is closed under the action of the gradient operator

then the corresponding factorization problem can always be brought to a canonical one via certain space extension transformations.

4. SOLUTION OF THE CANONICAL FACTORIZATION PROBLEM

Let us consider again the canonical problem as follows

$$Q = \exp\left\{t \sum_{j=1}^{N+1} \lambda \eta_j \frac{\partial}{\partial \eta_j} + t \sum_{j=1}^{N+1} \sum_{k=1}^{N+1} \sum_{l=1}^{N+1} b_{jkl} \eta_j \eta_k \frac{\partial}{\partial \eta_l}\right\} \quad (4.1)$$

$$Q_L = \exp\left\{t \lambda \sum_{j=1}^{N+1} \eta_j \frac{\partial}{\partial \eta_j}\right\} \quad (4.2)$$

$$Q\eta_1 = \{Q_L \left\{ \prod_{k=0}^{\infty} e^{\mu_k \eta_1^k \frac{\partial}{\partial \eta_1}} \right\} \eta_1\}_{\eta=\eta^{(f)}} \quad (4.3)$$

$$\eta_1^{(f)} = 1 \quad \eta_k^{(f)} = 0 \quad k = 2, 3, \dots, N+1 \quad (4.4)$$

Now, we can write the following formula for Q_L due to its special structure

$$Q_L = \prod_{j=1}^{N+1} e^{t \lambda \eta_j \frac{\partial}{\partial \eta_j}} \quad (4.5)$$

where the all factors except the leftmost one for $j = 1$ has a vanishing action on the remainder of the right side of Eq(4.3) due to the fact that every differentiation with respect to η_k -coordinates is followed by the multiplication with the corresponding η_k -coordinate so it is actionless on the factorization point for $k = 2, 3, \dots, N+1$. The μ -parameters can be altered with respect to their values at the factorization point due to the lack of derivatives with respect to $\eta_2, \eta_3, \dots, \eta_{N+1}$. Hence, we can write the following new form of the factorization formula

$$Q\eta_1 = \left\{ \prod_{k=0}^{\infty} e^{\sigma_k \eta_1^k \frac{\partial}{\partial \eta_1}} \eta_1 \right\}_{\eta=\eta^{(f)}} \quad (4.6)$$

where

$$\sigma_k = \mu_k(0, 0, \dots, 0, t) + \lambda t \delta_{k1} \quad k = 0, 1, \dots \quad (4.7)$$

Now, we shall deal with the partial differential equations of the μ -functions. We start with the equation for μ_0 ,

$$\frac{\partial \mu_0}{\partial t} = e^{-\lambda t} \{ \mu_0^2 (H_0 + D_{-1} \mu_0) - \mu_0 (H_1 + D_0 \mu_0) + H_2 + D_1 \mu_0 \} \quad \{\mu_0\}_{t=0} = 0 \quad (4.8a, b)$$

where H s and D s represents the following homogeneous functions and homogeneous operators respectively, the degree of each one is denoted by its subscript

$$H_0 = b_{111} \quad H_1 = \sum_{k=2}^{N+1} (b_{11k} + b_{1k1}) \eta_k \quad H_2 = \sum_{k=2}^{N+1} \sum_{l=2}^{N+1} b_{jkl} \eta_k \eta_l \quad (4.9a, b, c)$$

$$D_{-1} = \sum_{j=2}^{N+1} b_{j11} \frac{\partial}{\partial \eta_j} \quad D_0 = \sum_{j=2}^{N+1} \sum_{k=2}^{N+1} (b_{j1k} + b_{jk1}) \eta_k \frac{\partial}{\partial \eta_j} \quad (4.10a, b)$$

$$D_1 = \sum_{j=2}^{N+1} \sum_{k=2}^{N+1} \sum_{l=2}^{N+1} b_{jkl} \eta_k \eta_l \frac{\partial}{\partial \eta_j} \quad (4.10c)$$

If we define the following new variable instead of t , we can remove the explicit dependence on time.

$$\tau = \frac{1 - e^{\lambda t}}{\lambda} \quad (4.11)$$

Therefore,

$$\frac{\partial \mu_0}{\partial \tau} = \mu_0^2 (H_0 + D_{-1} \mu_0) - \mu_0 (H_1 + D_0 \mu_0) + H_2 + D_1 \mu_0 \quad \{\mu_0\}_{\tau=0} = 0 \quad (4.12)$$

This, however, implies that

$$\mu_0 = \sum_{k=0}^{\infty} \mu_0^{(k)} \tau^{k+1} \quad (4.13)$$

and the $\mu_0^{(k)}$ -coefficients satisfy the following recursion relation

$$(k+1) \mu_0^{(k)} = H_0 \sum_{l=0}^{k-2} \mu_0^{(k-l-2)} \mu_0^{(l)} + \sum_{l=0}^{k-3} \sum_{m=0}^{k-l-3} \mu_0^{(k-l-m-3)} \mu_0^{(m)} D_{-1} \mu_0^{(l)} - H_1 \mu_0^{(k-1)} - \sum_{l=0}^{k-2} \mu_0^{(k-l-2)} D_0 \mu_0^{(l)} + H_2 \delta_{k0} + D_1 \mu_0^{(k-1)} \quad k = 0, 1, \dots \quad (4.14)$$

where any μ -value with a negative superscript and any sum with a negative upper index will be taken equal to zero by definition. The explicit expressions for the first three coefficients are given below.

$$\mu_0^{(0)} = H_2 \quad \mu_0^{(1)} = \frac{1}{2} (D_1 H_2 - H_1 H_2) \quad (4.15a, b)$$

$$\mu_0^{(2)} = \frac{1}{3} (H_0 H_2^2 - H_1 D_1 H_2 - H_2 D_0 H_2) + \frac{1}{6} (H_1^2 H_2 + D_1^2 - H_2 D_1 H_1) \quad (4.15c)$$

The other coefficients can be evaluated in the same fashion. Although we are not going to explicitly present them, they have an important common property which is useful in the construction of the σ -coefficients and can be proved by means of the mathematical induction. We may express, $\mu_0^{(k)}$ as a homogeneous multinomial of $\eta_2, \eta_3, \dots, \eta_{N+1}$, the degree of which is equal to $k+2$. Hence, all the $\mu_0^{(k)}$ -coefficients vanish at the factorization point and this enables us to write

$$\sigma_0(t) = \mu_0(0, \dots, 0, t) = 0 \quad (4.16)$$

We may write the following partial differential equation for μ_1 which can be obtained after some intermediate steps

$$\frac{\partial \mu_1}{\partial \tau} - \mu_0^2 D_{-1} \mu_1 + \mu_0 D_0 \mu_1 - D_1 \mu_1 = 2\mu_0 H_0 - H_1 + 2\mu_0 D_{-1} \mu_0 - D_0 \mu_0 \quad \{\mu_1\}_{\tau=0} = 0 \quad (4.17)$$

The solution of this equation can be expressed as follows

$$\mu_1 = \sum_{k=0}^{\infty} \mu_1^{(k)} \tau^{k+1} \quad (4.18)$$

where the $\mu_1^{(i)}$ -coefficients can be determined with the aid of the following recursion

$$(k+1)\mu_1^{(k)} = \sum_{l=0}^{k-3} \sum_{m=0}^{k-l-3} \mu_0^{(k-l-m-3)} \mu_0^{(m)} D_{-1} \mu_1^{(l)} - \sum_{l=0}^{k-2} \mu_0^{(k-l-2)} D_0 \mu_0^{(l)} + D_1 \mu_1^{(k-1)} + 2H_0 \mu_0(k-1) - H_1 \delta_{k0} + 2 \sum_{l=0}^{k-2} \mu_0^{(k-l-2)} D_{-1} \mu_0^{(l)} - D_0 \mu_0^{(k-1)} \quad k = 0, 1, \dots \quad (4.19)$$

where the symbols which have negative upper indices are assumed to be zero as before.

The first three of these coefficients are given below

$$\mu_1^{(0)} = -H_1 \quad \mu_1^{(1)} = H_0 H_2 - \frac{1}{2} D_1 H_1 \quad (4.20a, b)$$

$$\mu_1^{(2)} = \frac{1}{3} (H_2 D_0 H_1 + H_2 D_1 H_0 - H_0 H_1 H_2) + \frac{2}{3} (H_0 D_1 H_2 + H_2 D_{-1} H_2) - \frac{1}{6} D_1^2 H_1 \quad (4.20c)$$

As can be shown via mathematical induction $\mu_1^{(i)}$ is a homogeneous multinomial, the degree of which is equal to $k+1$. Hence they vanish on the factorization point so we can write

$$\sigma_1(t) = \lambda t + \mu_1(0, \dots, 0, t) = \lambda t \quad (4.21)$$

With a little further effort the following form of the partial differential equation for μ_2 can be obtained.

$$\frac{\partial \mu_2}{\partial \tau} - \mu_0^2 D_{-1} \mu_2 + \mu_0 D_0 \mu_2 - D_1 \mu_2 = e^{-\mu_1} (2\mu_0 D_{-1} \mu_1 - D_0 \mu_1 - H_0 - D_{-1} \mu_0) \quad \{\mu_2\}_{\tau=0} = 0 \quad (4.22a, b)$$

This and the remaining equations for the other μ -functions can also be solved via series expansion in powers of τ and the corresponding σ -coefficients can be evaluated in the same fashion. We give only the first two of them by skipping the intermediate algebraic steps

$$\sigma_2(t) = b_{111} \tau \quad (4.23)$$

$$\sigma_3(t) = - \sum_{k=2}^{N+1} b_{k11} (b_{1k1} + b_{11k}) \tau \quad (4.24)$$

Therefore, we are able to evaluate the σ -coefficients for the canonical problem through a finite step algorithm. This can also be programmed for computational purposes. However, the construction of such a program up to any desired order within the limitations of computers is a quite delicate job. This will be a task for future work.

Since we now assume that the σ -coefficients, the Generators, can be evaluated, the final stage of the development is resulting action of LAEEO on the other coordinates $\eta_2, \eta_3, \dots, \eta_{N+1}$. Until now, we have dealt with the evaluation of $Q\eta_1$. However the undetermined ν -parameters give a certain degree flexibility in the scheme. Now, we can utilize these parameters to calculate the other terms like $Q\eta_2, Q\eta_3, \dots, Q\eta_{N+1}$. We start with the following identity which is satisfied by the transformation matrix, \bar{T} , of previous section

$$\bar{T}(\nu_1 + \bar{\nu}_1, \nu_2 + \bar{\nu}_2, \dots, \nu_{N+1} + \bar{\nu}_{N+1}) = \bar{T}(\nu_1, \nu_2, \dots, \nu_{N+1}) \bar{T}(\bar{\nu}_1, \bar{\nu}_2, \dots, \bar{\nu}_{N+1}) \quad (4.25)$$

Then we can write the following equations after a careful examination of the structure of $\bar{\xi}$ -approximants

$$\bar{\xi}^{(k)}(\nu_1, \nu_2, \dots, \nu_{N+1}; t) = Q\eta_k \quad k = 1, 2, \dots, N+1 \quad (4.26)$$

$$\begin{aligned} \bar{\xi}^{(1)}(\nu_1 + \bar{\nu}_1, \nu_2 + \bar{\nu}_2, \dots, \nu_{N+1} + \bar{\nu}_{N+1}; t) &= \bar{\xi}^{(1)}(\nu_1, \nu_2, \dots, \nu_{N+1}; t) + \\ &\sum_{l=1}^N \bar{\nu}_l \bar{\xi}^{(l+1)}(\nu_1, \nu_2, \dots, \nu_{N+1}; t) \end{aligned} \quad (4.27a)$$

$$\bar{\xi}^{(k)}(\nu_1 + \bar{\nu}_1, \nu_2 + \bar{\nu}_2, \dots, \nu_{N+1} + \bar{\nu}_{N+1}; t) = \bar{\xi}^{(k)}(\nu_1, \nu_2, \dots, \nu_{N+1}; t) \quad k = 2, \dots, N+1 \quad (4.27b)$$

The first equation above can be rewritten for $N+1$ different $\bar{\nu}$ -values such that the resulting set of linear equations can be solved for ξ -values appearing at its right hand side. Since the $\bar{\nu}$ -parameters can be considered as the elements of a vector lying in an N -dimensional space, we are able to choose N linearly independent vectors in this space, the elements of which correspond to the desired $\bar{\nu}$ -values. Hence the inversion mentioned above is always possible and the actions of LAEEO on the other coordinates are calculable. We call the λ and ν parameters as the "Characteristic Parameters of the Factorization" or simply "Characteristic Parameters". Their meaning will be clarified in a simple illustrative problem in the next section.

5. ROLES OF THE CHARACTERISTIC PARAMETERS IN THE FACTORIZATION SCHEME AND SIMPLE EXAMPLES.

In this section we deal with two simple examples to facilitate the explanation of our scheme. This will give insight into the concept of space extension and into the roles of the characteristic parameters in the factorization scheme. We do not give explicit computations, since substantiating results have already been given in our recent work on the one variable case [58,59]. The convergence theorems given in the accompanying paper are sufficient toward this end. We chose two typical examples. The purpose of the first one is directed at an explanation of space extension concept. The second one, however, reveals the importance of the characteristic parameters.

FIRST EXAMPLE:

This example is taken from the celestial mechanics. The motion of two particles interacting gravitationally can be expressed by the following differential equations (Hamilton's equations) and the accompanying initial conditions.

$$\frac{dx_j}{dt} = \frac{x_{j+6}}{m_1} \quad j = 1, 2, 3 \quad \frac{dx_j}{dt} = \frac{x_{j+6}}{m_2} \quad j = 4, 5, 6 \quad (5.1a)$$

$$\frac{dx_{j+6}}{dt} = -m_1 m_2 g \frac{x_j - x_{j+3}}{r} \quad j = 1, 2, 3 \quad \frac{dx_{j+6}}{dt} = m_1 m_2 g \frac{x_j - x_{j+3}}{r} \quad j = 4, 5, 6 \quad (5.1b)$$

$$r = \sqrt{(x_1 - x_4)^2 + (x_2 - x_5)^2 + (x_3 - x_6)^2} \quad (5.2)$$

$$x_j(0) = \alpha_j \quad j = 1, 2, \dots, 12 \quad (5.3)$$

The solution of these equations can be given through a LAEEO as follows

$$x_j(t) = \{e^{tL} z_j\}_{z=\alpha} \quad 1 \leq j \leq 12 \quad (5.4)$$

where

$$L = \sum_{j=1}^{12} f_j(z) \frac{\partial}{\partial z_j} \quad (5.5)$$

The descriptive functions of this system are the expressions on the right hand sides of the differential equations above. This problem is, of course, exactly soluble and our purpose here is one of illustrating the methodology.

As is very well-known, the first thing to do in attempting to solve the two body problem is the separation of the center of mass coordinates. Hence we also proceed in a similar way as follows

$$y_j = \frac{m_j x_j + m_2 x_{j+3}}{m_1 + m_2} \quad j = 1, 2, 3 \quad y_{j+3} = x_j - x_{j+3} \quad j = 1, 2, 3 \quad (5.6a)$$

$$y_{j+6} = x_{j+6} + x_{j+9} \quad j = 1, 2, 3 \quad y_{j+9} = \frac{m_2 x_{j+6} - m_1 x_{j+9}}{m_1 + m_2} \quad j = 1, 2, 3 \quad (5.6b)$$

$$L_C = \frac{1}{m_1 + m_2} (y_7 \frac{\partial}{\partial y_1} + y_8 \frac{\partial}{\partial y_2} + y_9 \frac{\partial}{\partial y_3}) \quad (5.7)$$

$$L_R = (\frac{1}{m_1} + \frac{1}{m_2}) (y_{10} \frac{\partial}{\partial y_4} + y_{11} \frac{\partial}{\partial y_5} + y_{12} \frac{\partial}{\partial y_6}) - \frac{m_1 m_2 g}{r^3} (y_4 \frac{\partial}{\partial y_{10}} + y_5 \frac{\partial}{\partial y_{11}} + y_6 \frac{\partial}{\partial y_{12}}) \quad (5.8)$$

$$L = L_C + L_R \quad (5.9)$$

Therefore, we obtain the Lie operator as a sum of two commutative operators: L_C corresponds to the motion of the center of mass, and L_R represents the relative motion of one particle with respect to other. The commutativity of these operators corresponds to the separation of variables. Indeed if we write

$$Q_C = e^{tL_C}, \quad Q_R = e^{tL_R}, \quad Q = Q_C Q_R \quad (5.9a, b, c)$$

then we can easily show that only one of Q_C and Q_R can create an evolution when we apply Q on one of the y -variables. In this sense, the change of coordinates from the z -variables to the y -variables can be considered a space contraction, because we have two separate systems with each of lower dimensions. The evolution characterized by Q_C is just a translation in space and has no more interesting feature for our purposes. On the other hand, it is useful increase the number of variables in the relative motion as

$$u_j = y_{j+3} \quad u_{j+3} = y_{j+9} \quad j = 1, 2, 3 \quad (5.10a, b)$$

$$u_7 = (y_4^2 + y_5^2 + y_6^2)^{-1/2} \quad (5.10c)$$

This enables us to remove the root structure appearing in L_R because of r as follows

$$L_R = \left(\frac{1}{m_1} + \frac{1}{m_2} \right) \left(u_4 \frac{\partial}{\partial u_1} + u_5 \frac{\partial}{\partial u_2} + u_6 \frac{\partial}{\partial u_3} \right) - m_1 m_2 g u_7^3 \left(u_1 \frac{\partial}{\partial u_4} + u_2 \frac{\partial}{\partial u_5} + u_3 \frac{\partial}{\partial u_6} \right) - \left(\frac{1}{m_1} + \frac{1}{m_2} \right) u_7^3 (u_1 u_4 + u_2 u_5 + u_3 u_6) \frac{\partial}{\partial u_7} \quad (5.11)$$

Now, we have descriptive functions which are multinomials in new variables. However their degree is greater than two, so we have to use further space extension, in other words, to increase the number of variables. To construct a rule of thumb, we emphasize that each new created variable to extend the space adds a new descriptive function such that it results from the action of L_R on the function which relates the new variable to the old ones. Hence, we choose a function appearing in the original descriptive functions as a new variable and check the effect of L_R on it. If it and the present descriptive functions are quadratically expressible in the new coordinates, then our goal is achieved, otherwise we can continue to create new variables until the quadratic structure is obtained. In the present case, it is reasonable to start with the most complicated term which can be considered as the product of u_7 and $u_7^2(u_1 u_4 + u_2 u_5 + u_3 u_6)$. Hence

$$L_R(\phi_1) = \left(\frac{1}{m_1} + \frac{1}{m_2} \right) (\phi_2 - 2\phi_1^2) - m_1 m_2 g \phi_3 \quad (5.12)$$

where

$$\phi_1 = u_7^2(u_1 u_4 + u_2 u_5 + u_3 u_6), \quad \phi_2 = u_7^2(u_4^2 + u_5^2 + u_6^2), \quad \phi_3 = u_7^5(u_1^2 + u_2^2 + u_3^2) = u_7^3 \quad (5.13a, b, c)$$

In Eq(5.13c) we have used the relation between u_7 and u_1, u_2, u_3 due to the fact that u_7 is an extended coordinate. Since the right side of Eq(5.12) and the present descriptive functions are quadratic in terms of the present variables and ϕ_1, ϕ_2, ϕ_3 , we can consider ϕ_2 and ϕ_3 as new variables in addition to ϕ_1 . However, the structures of $L_R(\phi_2)$ and $L_R(\phi_3)$ must be quadratic in all variables including the new ones for this purpose. Indeed, the following equalities verify this point

$$L_R(\phi_2) = -2\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\phi_1\phi_2 - 2m_1m_2g\phi_1\phi_3 \quad (5.14)$$

$$L_R(\phi_3) = -3\left(\frac{1}{m_1} + \frac{1}{m_2}\right)\phi_1\phi_3 \quad (5.15)$$

Hence, we can define the following new variables and extend our 7-dimensional space to a 10-dimensional one.

$$w_j = u_j \quad j = 1, \dots, 7 \quad w_{j+7} = \phi_j \quad j = 1, 2, 3 \quad (5.16a, b)$$

The Lie operator of the system in this space is as follows

$$L_R = \left(\frac{1}{m_1} + \frac{1}{m_2}\right)L_R^{(1)} - m_1m_2gL_R^{(2)} \quad (5.17)$$

$$L_R^{(1)} = w_4 \frac{\partial}{\partial w_1} + w_5 \frac{\partial}{\partial w_2} + w_6 \frac{\partial}{\partial w_3} - w_7 w_8 \frac{\partial}{\partial w_7} + (w_9 - 2w_8^2) \frac{\partial}{\partial w_8} - 2w_8 w_9 \frac{\partial}{\partial w_9} - 3w_8 w_{10} \frac{\partial}{\partial w_{10}} \quad (5.18)$$

$$L_R^{(2)} = w_1 w_{10} \frac{\partial}{\partial w_4} + w_2 w_{10} \frac{\partial}{\partial w_5} + w_3 w_{10} \frac{\partial}{\partial w_6} - w_{10} \frac{\partial}{\partial w_8} - 2w_8 w_{10} \frac{\partial}{\partial w_9} \quad (5.19)$$

Now, the remaining steps to arrive at the canonical problem are quite straightforward and we do not deal with them.

SECOND EXAMPLE:

Our second example is a linear differential equation system accompanied by given initial values as follows

$$\frac{dx_j}{dt} = \sum_{k=1}^N a_{jk} x_k \quad x_j(0) = \alpha_j \quad j = 1, \dots, N \quad (5.20a, b)$$

The corresponding canonical problem can be expressed as

$$y(t) = \{Q\eta_1\}_{\eta_1=0} \quad (5.21)$$

where

$$Q = e^{tL} \quad (5.22)$$

and

$$L = \lambda \sum_{j=1}^{N+1} \eta_j \frac{\partial}{\partial \eta_j} + \left(\sum_{l=1}^{N+1} T_{N+1,l} \eta_l \right) \sum_{j=1}^{N+1} \sum_{k=1}^{N+1} T_{jm}^{-1} \gamma_{mn} T_{mk} T_{nl} \quad (5.23)$$

where

$$\gamma_{jk} = (1 - \delta_{j, N+1})(\lambda \delta_{jk} - a_{jk})(1 - \delta_{k, N+1}) \quad (5.24)$$

and T_{jk}^{-1} denotes the element of the inverse of \mathbf{T} . If we, now, define

$$L_0 = \eta_1 \frac{\partial}{\partial \eta_1} \quad (5.25)$$

then we can show that the every commutator of L_0 with the remainder of L has an additive contribution which is proportional to the upperleftmost element of the matrix $\mathbf{T}^{-1}(\mathbf{A} - \lambda \mathbf{I})\mathbf{T}$. Therefore, if the matrix $\bar{\mathbf{W}}$ diagonalizes \mathbf{A} and λ is an eigenvalue which corresponding eigenvector is the first column of $\bar{\mathbf{W}}$, then we can take all the ν -parameters equal to zero, and furthermore all contributions of the commutators vanish and L_0 characterizes the total evolution. Hence, λ characterizes one of the modes of the system under consideration. On the other hand, if $\bar{\mathbf{W}}$ does not diagonalize \mathbf{A} then we can use the ν -parameters to make the first column of \mathbf{T} an eigenvector of \mathbf{A} . Therefore, the ν -parameters correspond to eigenvectors. They make possible the calculation not only the evolution of η_1 but the all remaining ones.

Although λ corresponds to the eigenvalues of \mathbf{A} , we do not have to assign the exact value to it. This may give certain advantages when the calculation of the eigenvalues becomes a cumbersome process. Hence we can use the factorization scheme even for approximating the exponential matrix. This subject is worthy of future study.

6. CONCLUDING REMARKS

In this work, we believe that an important step in the factorization of Lie-algebraic exponential evolution operators has been taken. A complete scheme was constructed for the multivariable LAEEO's. The effort was driven by a desire to create a method which is as simple as the one-variable case. The space extension techniques are used to produce the simplest factorization problem which has a special quadratic structure in the descriptive functions. However one has to be very careful about the use of the space extension concept. We assumed that the descriptive functions are infinitely differentiable, and this may not be the case and certain singularities may appear. Even in such cases the space extension may work as we have shown in the first example where r is identified obviously a singular structure. On the other hand, in the case of jump discontinuities, scheme may need further modification. The space of the coordinates may be separated into distinct regions, and a different space extension can be used for each region. Evidently, a regional factorization becomes necessary.

The convergence theorems are given in a companion paper. They are constructed for certain regions around the origin of an N -tuple space of the η -variables. The convergence for the entire N -tuple space is intensively studied.

The programming of the evaluation of the σ -variables is another interesting subject. Its foundations are presented in this work. However the construction of programs requires that sufficient attention be paid to the unusual structure of recursion relations. Symbolic programming languages like MACSYMA and REDUCE may be useful for generating executable codes.

Finally we believe that the presented scheme shows promise for being a powerful means for treating many application in science and engineering. Multi-dimensional problems are clearly the most interesting for study. One example attractive for study is the well known three body problem. This topic will be the focus of future work.

ACKNOWLEDGEMENT

The authors would like to thank Professor Hilmi Demiray for helpful comments.

REFERENCES

- [1] A. J. Dragt, "Lectures on Nonlinear Orbit Dynamics",
Physics of High Energy Particle Accelerators, *AIP Conference Proceedings* No 87,
(ed. by R. A. Carrigan et al.), New York, (1982)
- [2] A. J. Dragt and J. M. Finn, *J. Math. Phys.*, **17**, 2215, (1976)
- [3] A. J. Dragt and E. Forest, *J. Math. Phys.*, **24**, 2734, (1983)
- [4] M. Demiralp, *Bull. Tech. Univ. Istanbul*, **37**, 425, (1984)
- [5] R. K. Pathria, "Statistical Mechanics", *Pergamon Press*, New York, (1972)
- [6] D. Ruelle, "Statistical Mechanics", *W. A. Benjamin, Inc.*, New York, (1969)
- [7] R. Jancei, "Foundations of Classical and Quantum Statistical Mechanics",
Pergamon Press, Oxford, (1970)
- [8] C. Kittel, "Quantum Theory of Solids",
John Wiley & Sons, Inc., New York, (1963)
- [9] K. Huang, "Statistical Mechanics", *John Wiley & Sons, Inc.*,
New York, (1963)
- [10] L. D. Landau and E. M. Lifschitz, "Statistical Physics"
Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, (1958)
- [11] J. Schwinger, L. C. Biedenharn, and H. van Dam, (Eds),
"Quantum Theory of Angular Momentum", *Academic Press Inc.*, New York, (1965)
- [12] A. A. Abrikosov, L. P. Gor'kov and I. Ye Dzyaloshinskii,
"Quantum Field Theoretical Methods in Statistical Physics",
Pergamon Press, Oxford, (1965)
- [13] I. Prigogine, "Non-equilibrium Statistical Mechanics",
Interscience Publishers, New York, (1966)
- [14] E. Montroll, in "Lectures in Theoretical Physics", W. Downs and J. Downs (Eds),
Vol. III *Interscience Publishers, Inc.*, New York, (1961)
- [15] J. E. Campbell, *sl Proc. London Math. Soc.* **29**, 14, (1898)
- [16] H. F. Baker, *ibid sl Proc. London Math. Soc.* **34**, 347, (1902); **35**, 333, (1903);
2, 293, (1904); **3**, 24, (1904)
- [17] G. H. Weiss and A. A. Maraduduin, *J. Math. Phys.* **3**, 771, (1962)

- [18] W. Magnus, *Commun. Pure Appl. Math. Phys.* **7**, 649, (1954)
- [19] R. Kubo, W. E. Brittin, and L. G. Dunham, (Eds) *Interscience Publishers Inc.*, New York, (1959)
- [20] R. M. Wilcox, *J. Math. Phys.* **8**, 962, (1967)
- [21] N. H. Mc Coy, *Proc. Math. Acad. Sci. U. S.* **18**, 674, (1932)
- [22] J. E. Moyal, *Proc. Cambridge Phil. Soc.* **45**, 99, (1949)
- [23] E. P. Wigner, *Phys. Rev.* **40**, 749, (1932)
- [24] C. L. Mehta, *J. Math. Phys.* **5**, 677, (1964)
- [25] H. Weyl, "The Theory of Groups and Quantum Mechanics" *E. P. Dutton & Co., Inc.*, New York, (1931)
- [26] R. A. Sack, *Phil. Mag.* **3**, 497, (1958)
- [27] F. Bloch, *Z. Physik* **74**, 295, (1932)
- [28] R. M. Wilcox, *J. Chem. Phys.* **45**, 3312, (1966)
- [29] A. C. Zemach and R. J. Glauber, *Phys. Rev.* **101**, 118, (1956)
- [30] A. A. Maradudin, E. W. Montroll and G. H. Weiss, *Solid State Phys. Suppl.* **3**, 239, (1963)
- [31] N. D. Mermin, *J. Math. Phys.* **7**, 1038, (1966)
- [32] R. J. Glauber, *Phys. Rev.* **131**, (2766), (1963)
- [33] W. H. Louisell, "Radiation and Noise in Quantum Electronics", *Mc Graw-Hill Book Company, Inc.*, New York, (1964)
- [34] F. Fer, *Bull. Classe Sci. Acad. Roy. Belg.* **44**, 818, (1958)
- [35] K. Kumar, *J. Math. Phys.* **6**, 1928, (1965)
- [36] J. Wei and E. Norman, *J. Math. Phys.* **4**, 575, (1963)
- [37] H. Heffner and W. H. Louisell, *J. Math. Phys.* **6**, 474, (1965)
- [38] N. H. Mc Coy, *Proc. Edinburgh Math. Soc.* **3**, 118, (1932)
- [39] W. O. Kermack and W. H. Mc Crea, *Proc. Edinburgh Math. Soc.* **2**, 220, (1931)
- [40] L. Cohen, *J. Math. Phys.* **7**, 244, (1966)
- [41] D. J. Morgan and P. T. Landsberg, *Proc. Phys. Soc. (London)*, **86**, 261, (1965)
- [42] R. A. Cowley, *Adv. Phys.* **12**, 421, (1963)
- [43] R. M. Wilcox, *Phys. Rev.* **139**, A1281, (1965)
- [44] R. L. Peterson, *Rev. Mod. Phys.* **39**, 69, (1967)

- [45] R. Karplus and J. Schwinger, *Phys. Rev.* **73**, 1025, (1948)
- [46] R. F. Snider, *J. Math. Phys.* **5**, 1586, (1964)
- [47] R. Hermann, "Lie Groups for Physicists", W. A. Benjamin, Inc., New York, (1966)
- [48] W. Miller, Jr., "Symmetry Groups and Their Applications",
Academic Press, New York, (1972)
- [49] G. Hochschild, "The Structure of Lie Groups", *Holden-Day, Inc.*, (1965)
- [50] C. Von Westenholz, "Differential Forms in Mathematical Physics",
North-Holland Publishing Company, New York, (1981)
- [51] B. F. Schutz, "Geometrical Methods of Mathematical Physics",
Cambridge University Press, London, (1980)
- [52] S. Helgason, "Differential Geometry and Symmetric Spaces",
Academic, New York, (1962)
- [53] M. Hausner and J. T. Schwartz, "Lie Groups; Lie Algebras",
Gordon and Breach, New York, (1968)
- [54] W. Gröbner, "Die Lie-Reihen und Ihre Anwendungen",
veb Deutscher Verlag der Wissenschaften, Berlin, (1967)
- [55] C. Wulfman and H. Rabitz, *J. Phys. Chem.*, **90**, 2264, (1986)
- [56] L. M. Hubbard, C. Wulfman and H. Rabitz, *J. Phys. Chem.*, **90**, 2273, (1986)
- [57] R. L. Anderson, J. Harnad and P. Winternitz, *Physica* **4D**, 164, (1982)
- [58] M. Demiralp, H. Rabitz, "Factorization of certain evolution operators using Lie
algebra: Formulation of the method" (*to be published*)
- [59] M. Demiralp, H. Rabitz, "Factorization of certain evolution operators using Lie
algebra: Convergence theorems (*to be published*)

Appendix M

13. Global Sensitivity Analysis of Nonlinear Chemical Kinetic Equations Using Lie Groups; I. Determination of One-parameter Groups, C.E. Wulfman and H. Rabitz, J. Math. Chem., 3, 243 (1989).

GLOBAL SENSITIVITY ANALYSIS OF NONLINEAR CHEMICAL KINETIC EQUATIONS USING LIE GROUPS: I. DETERMINATION OF ONE-PARAMETER GROUPS

C.E. WULFMAN

Department of Physics, The University of the Pacific, Stockton, CA 95207, USA

and

H. RABITZ

Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

Received 14 December 1987
(in final form 2 January 1989)

Abstract

We introduce one-parameter groups of transformations that effect wide-ranging changes in the rate constants and input/output fluxes of homogeneous chemical reactions involving an arbitrary number of species in reactions of zero, first and second order. Each one-parameter group is required to convert every solution of such elementary rate equations into corresponding solutions of a one parameter family of altered elementary rate equations. The generators of all allowed one-parameter groups are obtained for systems with N species using an algorithm which exactly determines their action on the rate constants, and either exactly determines or systematically approximates their action on the concentrations. Compounding the one-parameter groups yields all many-parameter groups of smooth time-independent transformations that interconvert elementary rate equations and their solutions.

1. Introduction

The response of kinetic systems over extensive regions of their physical parameter space – the space of rate constants and input/output fluxes – is of wide interest in many different contexts. For example, chemical system modelling can involve solving large numbers of coupled rate equations with considerable uncertainties in many values of the rate constants. In other problems some of the system parameters (e.g. input fluxes of chemical species) may actually be controlled, but determining the optimum choice of parameter values would require exploring a large domain of

control-parameter space. Conventional gradient-based local sensitivity analysis techniques [1] have limited applicability in problems of this type. In addition, fully statistically-based approaches [2] do not allow for an analysis of the structure of the parameter space. Other methodologies [3] based on repeated sampling of points in the parameter space suffer from the same problem and often require an impractical amount of computational labor.

In two previous papers, an alternative approach to sensitivity analysis, using Lie transformation groups, was introduced as a method for investigating the consequences of large changes in parameters in kinetic equations [4,5]. The present paper extends this effort into the realm of nonlinear kinetics.

The thrust of this work is the development of a systematic procedure that yields mappings which transform solutions of a system of kinetic equations through the hyperdimensional space defined by all rate constants, chemical species, and time. Here we will not, however, consider transformations of the time variable. We also do not allow the transformed rate constants to be explicit functions of the concentration variables.

The mappings are achieved by the application of operators $T(a) = \exp(aU)$ of one-parameter groups, where a is a real parameter and U is a group generator of Lie type. This generator is a first-order differential operator which may act on all physical parameters and variables of the kinetic system. Symbolizing concentrations by x_i and rate constants by k_μ , the generator here takes the form

$$U = \sum h_i(x, k) \partial / \partial x_i + \sum g_\mu(k) \partial / \partial k_\mu. \quad (1.1)$$

Here, x represents the set of x_i and k represents the set of k_μ . Henceforth, x, k represent vectors with components x_i and k_μ in a Euclidean space of x, k . The operator of finite transformations $T(a) = \exp(aU)$ acts as follows:

On a rate constant k_μ :

$$T(a)k_\mu = \bar{k}_\mu = K_\mu(k; a); K_\mu(k; 0) = k_\mu. \quad (1.2a)$$

On the concentration x_i of species i :

$$T(a)x_i = \bar{x}_i = X_i(x, k; a); X_i(x, k; 0) = x_i. \quad (1.2b)$$

Figure 1.1 depicts the type of mapping being considered.

As indicated in eqs. (1.2), assigning the group parameter a the value zero gives the identity transformation. As a is shifted from zero by infinitesimal and then finite amounts, changes in k and x develop which are at first infinitesimal, and then become increasingly profound. For a fixed value of the parameter a , $T(a)$ acts on the moving vector $x(t)$ to give the transformed vector $\bar{x}(t) = X(x(t), k, a)$. It thus transforms the curve in concentration space described by $x(t)$ into a new curve depicted

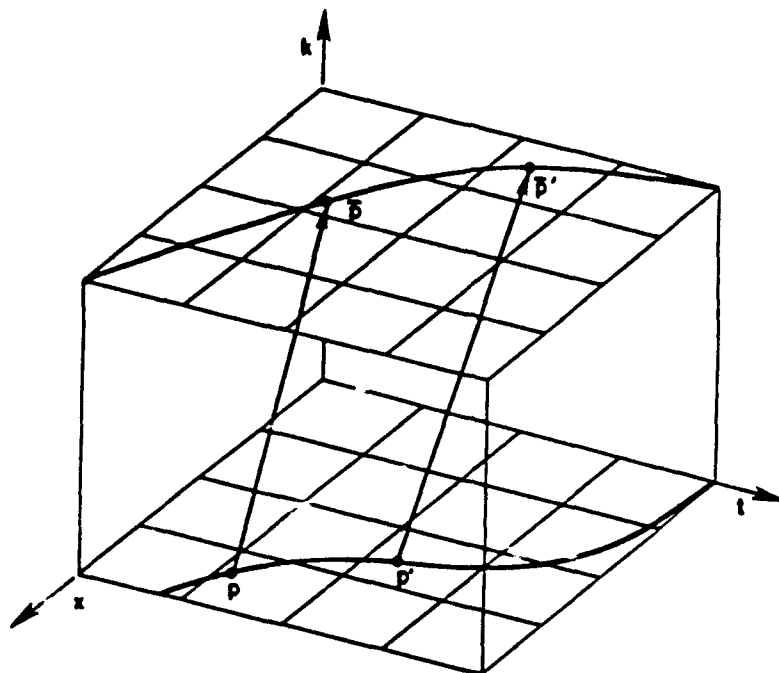


Fig. 1.1. The mappings in x, k, t space. The mappings $P \rightarrow \bar{P}$ represent the concentration changes $x \rightarrow \bar{x}$ and the changes in rate constants $k \rightarrow \bar{k}$, while the time t is held fixed. As \bar{k} is not a function of x or t , the trajectory $P \rightarrow P'$ is mapped into a trajectory $\bar{P} \rightarrow \bar{P}'$ that lies in a hyperplane of constant \bar{k} .

ing an altered evolution of chemical concentrations. By changing the value of the parameter a , one is able to convert an initial evolution curve into a one-parameter family of evolution curves. Thus, in fig. 1.1 the upper curve may be considered as one member of a family of transformed curves, a curve obtained by giving the group parameter a specific value. The value of the group parameter a can be assigned by the investigator, but it is neither a rate constant nor a concentration. Its chemical significance is determined by the functions K_i and X_μ in (1.2). This significance, and that of the generator U , can be assessed by investigating the action of the operator of the infinitesimal transformation $T(\delta a)$.

Letting $a \rightarrow \delta a$, one has

$$\exp(aU) \rightarrow \exp(\delta a U) \sim 1 + \delta a U. \quad (1.3)$$

Thus, for an infinitesimal transformation,

$$\bar{x}_i = x_i + \delta a U x_i = x_i + \delta a h_i(x, t); \quad \bar{k}_\mu = k_\mu + \delta a U k_\mu = k_\mu + \delta a g_\mu(k). \quad (1.4)$$

Consequently, if one defines δx_i as $\bar{x}_i - x_i$ and δk_μ as $\bar{k}_\mu - k_\mu$ in (1.4) one has

$$\delta x_i = \delta a h_i(x, k), \quad \delta k_\mu = \delta a g_\mu(k). \quad (1.5)$$

It follows that $T(\delta a)$ changes the concentration x_i by an amount $\delta a h_i$ that may depend upon all concentrations x and rate constants k . Similarly, the transformation changes the rate constant k_μ by an amount $\delta a g_\mu$ that may depend upon all rate constants k . As an example, consider the generator

$$U = k_{11} x_1 \partial / \partial x_1 + 2 \partial / \partial k_{10} \quad (1.6)$$

and its action on a system involving a single species obeying the rate equation

$$dx_1/dt = k_{10} + k_{11} x_1 + k_{111} x_1^2. \quad (1.7)$$

This generator determines a shift in the concentration x_1 by an amount $\delta x_1 = k_{11} x_1 \delta a$, i.e. a shift proportional to the product of the concentration and the second-order rate constant. This determines a consequent shift in dx_1/dt by an amount $d(k_{11} x_1 \delta a)/dt = \delta a k_{11} dx_1/dt$. It also determines a shift $\delta k_{10} = 2 \delta a$ in the flux k_{10} . The generator does not affect either k_{11} or k_{111} .

Now, if it were true that the shifted concentration obeyed the same rate equation with the shifted value of k_{10} , the generator (1.6) could be of use in investigations of the consequences of changing the rate of supply or removal of the reagent. The operator $T(a) = \exp(aU)$ could then be used to determine the relation between changes in the flux and changes in the concentration x , the extent of both changes being determined by the value of the parameter a . However, the U of (1.6) was chosen at random and can not be expected at each value of t to convert $x(t)$ into $\bar{x}(t)$ that obey the altered rate equation.

If the U of (1.6) had the property that $UF = 0$, where

$$F = (k_{10} + k_{11} x_1 + k_{111} x_1^2), \quad (1.8)$$

then $\exp(aU)$ acting on the right-hand side of (1.7) would leave it unchanged, i.e. not change the reaction rate. This is because

$$\exp(aU)F = (1 + aU + \frac{1}{2}aUaU + \dots)F \quad (1.9a)$$

would then give

$$F + 0 + 0 + \dots = F. \quad (1.9b)$$

This is not, however, the restriction we wish to impose.

The restrictions we impose upon the $T(a)$, and hence the U 's, so as to obtain chemical information from them are as follows: Each $T(a)$ will be required to have a unique action on all k, x , in an elementary kinetic equation, map contiguous values of k_μ and x_i into contiguous values of \bar{k}_μ and \bar{x}_i , and give \bar{k} and \bar{x} that also satisfy elementary kinetic equations (cf. section 2 below). In addition, we shall require that all the variables a, x, k are real. Taken together, these requirements ensure that the transformation $T(a)$ maps solutions of the set of kinetic equations

$$dx_i/dt = k_{i0} + k_{ij}x_j + k_{ijj'}x_jx_{j'} \quad (1.10a)$$

into solutions of the set of transformed equations

$$d\bar{x}_i/dt = \bar{k}_{i0} + \bar{k}_{ij}\bar{x}_j + \bar{k}_{ijj'}\bar{x}_j\bar{x}_{j'} \quad (1.10b)$$

They impose restrictions on the form of the generators U sufficient to ensure that the U may be determined algorithmically. Because of this, one has available a systematic method for investigating the manner in which changes in rate constants are associated with changes in species concentrations and their time evolution. These restrictions are not equivalent to requiring that $T(a)$ leave reaction rates dx_i/dt invariant.

In the next section, we outline an algorithm for determining the allowed Lie generators U and use it to completely determine the terms in the generators which govern the transformation of rate constants of kinetic systems with an arbitrary number of species. The remaining terms in the generators, governing the transformation of species concentrations, are approximated by power series whose zero-, first-, and second-order terms we determine.

2. Derivation of approximate invariance operators: Their action

Let a set of kinetic equations be given as

$$\dot{x} = r(x, k),$$

with

$$\dot{x} = dx/dt; \quad -\infty < t, x_i, \dot{x}_i < \infty$$

$$r = (r_1, r_2, \dots)$$

$$r_i = k_{i0} + k_{ij}x_j + k_{ijj'}x_jx_{j'}, \quad j' \geq j,$$

$$i, j, j' = 1, 2, \dots; \quad -\infty < k_\mu < \infty. \quad (2.1)$$

The evolution operator of this system is then $\exp(tV)$, with

$$V = r \cdot \nabla_x, \quad \nabla_x = (\partial/\partial x_1, \partial/\partial x_2, \dots). \quad (2.2a)$$

That is,

$$\bar{x} = \exp(tV)x = X(x, k; t) \quad (2.2b)$$

is the vector that x evolves into after a time interval t .

Define the operator $\exp(aU)$ of a one-parameter Lie group of transformations with real parameter a , $(-\infty < a < \infty)$ and generator U of the form

$$U = h \cdot \nabla_x + g \cdot \nabla_k,$$

where

$$h = (h_1, h_2, \dots), \quad h_i = h_{i0} + h_{ij}x_j + h_{ijj'}x_jx_{j'} + \dots$$

$$h_{ijj'} = h_{ij'j}, \text{ etc.} \quad (2.3)$$

$$g \cdot \nabla_k = \sum g_{im} \partial/\partial k_{im}, \quad m = 0, j, jj' \dots$$

Here, and in the remainder of the paper, we use the index m in h_{im} , k_{im} and g_{im} to signify any of the values $0, j, jj' \dots$

The coefficients h_{im} may in general be allowed to be explicit functions of t , x , k . The coefficients g_{im} are not allowed to depend upon x or t but can depend upon k . In ref. [4] it was shown that with these restrictions the action of $\exp(aU)$ on the variables x and k is to give a set of transformed variables \bar{x} and \bar{k} in which the \bar{k} have fixed values that do not change with time, while the \bar{x} are, like the x , running variables whose values change with time. On transformation, the new values of the k_{im} depend upon the old values, but not upon x or t : geometrically, the space of the k_{im} is an invariant subspace of the space of x, t, k . The k_{im} are allowed to take on any real values, and in particular may take on the special value zero without altering the general form of the equations given in (2.1).

It was also shown in ref. [4] that the transformed equations will be of the same general form, (2.1), with x replaced by \bar{x} and k replaced by \bar{k} if and only if

$$W \equiv [V, U] + \partial U/\partial t = 0. \quad (2.4)$$

In this paper, we shall require that the h_{im} are time independent so that here $\partial U/\partial t$ is zero. W is then easily seen to be of the form

$$W = w \cdot \nabla_x,$$

with

$$w = (w_1, w_2, \dots)$$

and

$$w_i = w_{i0} + w_{ij}x_j + w_{ijj'}x_jx_{j'} + \dots \quad (2.5)$$

For (2.4) to hold in the time-independent case, it is necessary that each of the coefficients w_{im} vanish identically. For reasons explained below, we shall at first approximate h by the terms explicitly listed in (2.3) and only require that the coefficients given explicitly in (2.5) vanish. The resulting quadratic approximation to the generators U will later be improved by methods discussed in the succeeding paper II. Each w_{im} in (2.4) is a bilinear function of the k_{im} and h_{im} , and is linear in the g_{im} . Our first problem is to determine the h_{im} and the g_{im} .

Before determining the generators in which h is quadratically approximated, it is helpful to understand the effect of allowing h to depend upon polynomials of arbitrary degree in the x_i . To this end, we classify the contributions to U, V, W according to their degree in x . We write

$$r = r^{(0)} + r^{(1)} + r^{(2)}, \quad (2.6)$$

where $r^{(p)}$ is a homogeneous polynomial of degree p in x , and we write

$$V^{(p-1)} = r^{(p)} \cdot \nabla_x$$

to indicate that the corresponding contribution to the generator is of one degree less. Then

$$\begin{aligned} V &= (r^{(0)} + r^{(1)} + r^{(2)}) \cdot \nabla_x = V^{(-1)} + V^{(0)} + V^{(1)} \\ U &= (h^{(0)} + h^{(1)} + h^{(2)} + h^{(3)} + \dots) \cdot \nabla_x + g \cdot \nabla_k \\ &= U^{(-1)} + U^{(0)} + U^{(1)} + U^{(2)} + \dots + g \cdot \nabla_k, \\ W &= [U, V] = W^{(-1)} + W^{(0)} + W^{(1)} + W^{(2)} + \dots \end{aligned} \quad (2.7)$$

Now the commutator of $U^{(m)}$ and $V^{(n)}$ is of degree $m+n$, and the commutator of $k \cdot \nabla_g$ and $V^{(n)}$ is of degree n . Thus, the vanishing of W requires that

$$0 = W^{(-1)} = [U^{(-1)}, V^{(0)}] + [U^{(0)}, V^{(-1)}] + g \cdot \nabla_k (r^{(0)} \cdot \nabla_x) \quad (2.8a)$$

$$0 = W^{(0)} = [U^{(-1)}, V^{(1)}] + [U^{(0)}, V^{(0)}] + [U^{(1)}, V^{(-1)}] + g \cdot \nabla_k (r^{(1)} \cdot \nabla_x) \quad (2.8b)$$

$$0 = W^{(1)} = [U^{(0)}, V^{(1)}] + [U^{(1)}, V^{(0)}] + [U^{(2)}, V^{(-1)}] + g \cdot \nabla_k (r^{(2)} \cdot \nabla_x) \quad (2.8c)$$

$$0 = W^{(2)} = [U^{(1)}, V^{(1)}] + [U^{(2)}, V^{(0)}] + [U^{(3)}, V^{(-1)}] \quad (2.8d)$$

$$0 = W^{(p)} = [U^{(p-1)}, V^{(1)}] + [U^{(p)}, V^{(0)}] + [U^{(p+1)}, V^{(-1)}], \quad p > 3. \quad (2.8e)$$

Note that each of these equations stands for a set of separate equations $w_{im} = 0$, where w_{im} is the coefficient of

$$\partial/\partial x_i, x_j \partial/\partial x_i, x_j x_j \partial/\partial x_i \dots \text{as } m = 0, j, jj' \dots \quad (2.9)$$

A key feature of the set of equations $w_{im} = 0$ is the fact that their rank is much less than their order, so that their solution contains many free parameters. If we do not allow cubic and higher degree polynomials in x into U and W , we find that the equations $w_{im} = 0$ for $W^{(-1)}, W^{(0)}, W^{(1)}$ are the set of simultaneous linear equations

$$\begin{aligned} \sum_p \{ h_{p0} k_{ip} - h_{ip} k_{p0} \} + g_{i0} &= 0, \quad i = 1, 2, \dots, n \\ \sum_p \{ h_{p0} (k_{ijp} + k_{ipj}) + h_{pj} k_{ip} - j_{ip} k_{pj} - (h_{ipj} + h_{ijp}) k_{p0} \} + g_{ij} &= 0 \\ i, j &= 1, 2, \dots, n \\ \sum_p \{ h_{pj} (k_{ipk} + k_{ikp}) + h_{pk} (k_{ipj} + k_{ijp}) - h_{ip} (k_{pjk} + k_{pkj}) + (h_{pj} + h_{pkj}) k_{ip} \\ &\quad - (h_{ipk} + h_{ikp}) k_{pj} - (h_{ipj} + h_{ijp}) k_{pk} \} + g_{ijk} + g_{ikj} = 0 \\ i, j, k &= 1, 2, \dots, n. \end{aligned} \quad (2.10)$$

In this "quadratic" approximation, each component of g is uniquely determined by a single equation if one chooses r to be a one-term homogeneous polynomial. Since the general solution of the equation is an arbitrary linear combination of these special solutions, one may make this choice without any loss of generality. In this linear combination, the coefficients may be arbitrary functions of the k_{im} . We shall say that the generators U_m in a collection are "independent" if no linear combination of them

$$\sum c_m U_m$$

is identically equal to zero when the coefficients c_m in the linear combination are not functions of x .

The remaining sections of this paper will make use of the quadratic approximation to the generators and the approximation to (2.8) obtained by dropping all $W^{(p)}$ with p greater than 1. We shall term this twofold approximation the "quadratic approximation". In paper II, we will investigate more accurate approximations to the generators and show that the quadratic approximation is of great utility.

In the two-species case, we obtain twelve equations $w_{im} = 0$ from the quadratic approximation to (2.8). Their general solution is a linear combination of twelve independent special solutions. Each special solution fixes a generator U , listed in table 2.1. The generators whose h 's are of zero or first order in x are exact solutions of (2.3).

Inspecting table 2.1, the reader will note that we have chosen the U_{im} to be of the form (here, g^{i0} is the g vector of U_{i0} , etc.)

$$\begin{aligned} U_{i0} &= \partial/\partial x_i + g^{i0} \cdot \nabla_k, & U_{ij} &= x_j \partial/\partial x_i + g^{ij} \cdot \nabla_k \\ U_{ijj'} &= x_j x_{j'} \partial/\partial x_i + g^{ijj'} \cdot \nabla_k. \end{aligned} \quad (2.11)$$

That is, eqs. (2.10) allow one to choose the action of each U upon the species concentrations and then determine the action on the kinetic coefficients that is required to leave the kinetic equations invariant up through terms quadratic in the concentrations.

This procedure generalizes to systems of three or more species. As a result, one can easily obtain analogously exact and quadratically approximated invariance generators U for kinetic systems (2.1) involving an arbitrary number of species. In the general case, the generators obtained with the aid of eqs. (2.10) are:

$$\begin{aligned} U_{i0} &= \partial/\partial x_i - \sum_j k_{ij} \partial/\partial k_{j0} - \sum_{m \neq i} k_{jmi} \partial/\partial k_{jm} - 2 \sum_j k_{jii} \partial/\partial k_{ji} \\ U_{ii} &= x_i \partial/\partial x_i + k_{i0} \partial/\partial k_{i0} + \sum_{j \neq i} k_{ij} \partial/\partial k_{ij} - k_{iii} \partial/\partial k_{iii} \\ &+ \sum_{j, m \neq i} k_{ijm} \partial/\partial k_{ijm} - \sum_{j \neq i} k_{ji} \partial/\partial k_{ji} \\ &- 2 \sum_{j \neq i} k_{jii} \partial/\partial k_{jii} - \sum_{m, j \neq i} k_{jim} \partial/\partial k_{jim}. \end{aligned}$$

For $j \neq i$:

Table 2.1
Generators of invariance transformations

Generator	h	g										
		h_1	h_2	g_{10}	g_{11}	g_{12}	g_{111}	g_{112}	g_{122}	g_{20}	g_{21}	g_{22}
U_{10}	1	0	0	$-k_{11}$	$-2k_{111}$	$-k_{112}$	0	0	0	$-k_{21}$	$-2k_{211}$	$-k_{212}$
U_{11}	x_1	0	0	k_{10}	0	k_{12}	$-k_{111}$	0	k_{122}	0	$-k_{21}$	0
U_{12}	x_2	0	0	k_{20}	k_{21}	$(k_{22} - k_{11})$	k_{211}	$(k_{212} - 2k_{111})$	$(k_{222} - k_{112})$	0	0	$-k_{212}$
U_{111}	x_1^2	0	0	0	$2k_{10}$	0	k_{11}	$2k_{12}$	0	0	0	0
U_{112}	$x_1 x_2$	0	0	0	k_{20}	k_{10}	k_{21}	k_{22}	k_{12}	0	0	0
U_{122}	x_2^2	0	0	0	0	$2k_{20}$	0	$2k_{21}$	$(2k_{22} - k_{11})$	0	0	$-k_{21}$
U_{20}	0	1	0	$-k_{12}$	$-k_{112}$	$-2k_{122}$	0	0	0	$-k_{22}$	$-k_{212}$	$-2k_{222}$
U_{21}	x_1	0	x_1	0	$-k_{12}$	0	$-k_{112}$	$-2k_{122}$	0	k_{10}	$(k_{11} - k_{22})$	k_{12}
U_{22}	x_2	0	x_2	0	0	$-k_{12}$	0	$-k_{112}$	$-2k_{122}$	k_{20}	k_{21}	0
U_{211}	x_1^2	0	x_1^2	0	0	0	$-k_{12}$	0	0	0	$(2k_{11} - k_{22})$	$2k_{12}$
U_{212}	$x_1 x_2$	0	$x_1 x_2$	0	0	0	0	$-k_{12}$	0	0	k_{21}	k_{11}
U_{222}	x_2^2	0	x_2^2	0	0	0	0	0	$-k_{12}$	0	0	$2k_{21}$

$$\begin{aligned}
U_{ij} = & x_j \partial / \partial x_i + k_{j0} \partial / \partial k_{i0} + k_{ji} \partial / \partial k_{ii} + (k_{jj} - k_{ii}) \partial / \partial k_{ij} \\
& + \sum_{m \neq i, j} k_{jm} \partial / \partial k_{im} - \sum_{m \neq i} k_{mi} \partial / \partial k_{mj} + k_{jii} \partial / \partial k_{iii} \\
& + (k_{jij} - 2k_{iii}) \partial / \partial k_{iij} + (k_{jjj} - k_{iij}) \partial / \partial k_{ijj} \\
& + \sum_{m \neq i, j} k_{jim} \partial / \partial k_{iim} + \sum_{m \neq i, j} (k_{jjm} - k_{iim}) \partial / \partial k_{ijm} \\
& - 2 \sum_{m \neq i} k_{mii} \partial / \partial k_{mij} - \sum_{m, n \neq i} k_{min} \partial / \partial k_{mjn} \\
U_{iii} = & x_i x_i \partial / \partial x_i + 2k_{i0} \partial / \partial k_{ii} + k_{ii} \partial / \partial k_{iii} \\
& + 2 \sum_{j \neq i} k_{ij} \partial / \partial k_{iij} - \sum_{j \neq i} k_{ji} \partial / \partial k_{jii} .
\end{aligned} \tag{2.12}$$

For $j \neq i$:

$$\begin{aligned}
U_{iij} = & x_i x_j \partial / \partial x_i + k_{j0} \partial / \partial k_{ii} + k_{i0} \partial / \partial k_{ij} + \sum k_{jm} \partial / \partial k_{iim} + 2k_{ij} \partial / \partial k_{iij} \\
& + \sum_{m \neq i, j} k_{im} \partial / \partial k_{ijm} - \sum_{m \neq i} k_{mi} \partial / \partial k_{mij} \\
U_{ijj} = & x_j x_j \partial / \partial x_i + 2k_{j0} \partial / \partial k_{ij} + 2k_{ji} \partial / \partial k_{iij} + (2k_{jj} - k_{ii}) \partial / \partial k_{ijj} \\
& + 2 \sum_{m \neq i, j} k_{jm} \partial / \partial k_{ijm} - \sum_{m \neq i} k_{mi} \partial / \partial k_{mjj} .
\end{aligned} \tag{2.13}$$

For i, j, j' all different:

$$\begin{aligned}
U_{ijj'} = & x_j x_{j'} \partial / \partial x_i + k_{j0} \partial / \partial k_{ij'} + k_{j'0} \partial / \partial k_{ij} + k_{jj'} \partial / \partial k_{ij'j'} \\
& + k_{j'j} \partial / \partial k_{iij} + \sum_{m \neq j'} k_{jm} \partial / \partial k_{imj'} \\
& + \sum_{m \neq j} k_{j'm} \partial / \partial k_{ijm} - \sum_m k_{mi} \partial / \partial k_{mjj'} .
\end{aligned} \tag{2.14}$$

In this list, the generators U_{i0} , U_{ii} , and U_{ij} exactly satisfy the determining eqs. (2.8). The generators U_{iii} , U_{iij} , and $U_{ijj'}$ satisfy (2.8) in quadratic approximation.

3. Finite transformations

As mentioned earlier, corresponding to each generator U there is an operator $\exp(aU)$ of finite transformations. One way to determine the effect of this upon each variable x_i , k_μ is to expand the exponential in powers of aU , carry out the indicated actions and sum the resulting series, which sometimes terminates, has evident recursiveness, or is recognizable as the MacLaurin expansion of a simple function. Often, a more practical method is to integrate the set of equations [4]:

$$\delta a = \frac{\delta x_1}{h_1} = \frac{\delta x_2}{h_2} \dots = \frac{\delta k_{10}}{g_{10}} = \frac{\delta k_{20}}{g_{20}} = \frac{\delta k_{11}}{g_{11}} \dots = \frac{\delta k_{2jj'}}{g_{2jj'}} \dots \quad (3.1)$$

When the h are of the form we have chosen, the necessary integrations can all be carried out analytically.

Note that the only concentration altered by $T_{im}(a) = \exp(aU_{im})$ is x_i . One finds using (3.1):

$$\begin{aligned} T_{i0}(a)x_i &= x_i + a, & T_{ii}(a)x_i &= x_i e^a, \\ T_{ij}(a)x_i &= x_i + ax_j, \quad j \neq i, & T_{iii}(a)x_i &= x_i/(1 - ax_i) \\ T_{ijj'}(a)x_i &= x_i e^{ax_j}, \quad j \neq i, & T_{ijj'}(a)x_i &= x_i + ax_j x_{j'}, \quad i \neq j, j'. \end{aligned} \quad (3.2)$$

The effect of each of the finite transformation operators on the kinetic parameters k_{im} are listed in table 2.2. As an example, one finds from table 2.2 that T_{10} acting on k_{10} gives $\bar{k}_{10} = k_{10} - ak_{11}$.

Because $T_{10}(a)$ and $T_{20}(b)$ leave \dot{x}_1 and \dot{x}_2 invariant, their action on the kinetic equations can be determined by replacing x_1 by $x_1 + a$, or x_2 by $x_2 + b$, in r and determining the coefficients $c_{ijj'}$ of the various powers $x_j x_{j'}$ of the concentrations in the equation for \dot{x}_i . Then one finds $\bar{k}_{ijj'} = c_{ijj'}$. Because T_{10} and T_{20} carry out translations of x while leaving the kinetic equations invariant in the generalized sense that the quadratic polynomial form of r is preserved, we shall term them "invariant translation" operators.

In all the generators other than the U_{10} , the operator $\partial/\partial x_i$ is premultiplied by either x_i or x_j . As a consequence, these generators vanish at the origin of x . Because of this, the corresponding operators of finite transformations T cannot move a point at the origin. If one lets U be a linear combination of the generators in table 2.1, the finite transformations may be obtained by solving eqs. (3.1) *de novo*.

Before concluding this paper, we would call attention to some geometrical properties of our transformations. First note that the evolution generator V is a special type of U with $g = 0$, and that the corresponding operator of finite trans-

Table 2.2(a)
Finite transformations of k

	\bar{k}_{10}	\bar{k}_{11}	\bar{k}_{12}	\bar{k}_{111}	\bar{k}_{112}	\bar{k}_{122}
T_{10}	$k_{10} - ak_{11} + a^2 k_{111}$	$k_{11} - 2ak_{111}$	$k_{12} - ak_{112}$	k_{111}	k_{112}	k_{122}
T_{11}	$e^a k_{10}$	k_{11}	$e^a k_{12}$	$e^{-a} k_{111}$	k_{112}	$e^a k_{122}$
T_{12}	$k_{10} + ak_{20}$	$k_{11} + ak_{21}$	$k_{12} - a^2 k_{21} + a(k_{22} - k_{11})$	$k_{111} + ak_{211}$	$k_{112} - 2a^2 k_{211} + a(k_{212} - 2k_{111})$	$k_{122} + a(k_{222} - k_{112}) + a^2(k_{111} - k_{212}) + a^3 k_{211}$
T_{111}	k_{10}	$k_{11} + 2ak_{10}$	k_{12}	$k_{111} + ak_{11} + a^2 k_{10}$	$k_{112} + 2ak_{12}$	k_{122}
T_{112}	k_{10}	$k_{11} + ak_{20}$	$k_{12} + ak_{10}$	$k_{111} + ak_{21}$	$k_{112} + ak_{22}$	$k_{122} + ak_{12} + \frac{1}{2}a^2 k_{10}$
T_{122}	k_{10}	k_{11}	$k_{12} + 2ak_{20}$	k_{111}	$k_{112} + 2ak_{21}$	$k_{122} + a(2k_{22} - k_{11})$

Table 2.2(b)
Finite transformations of k

	\bar{k}_{20}	\bar{k}_{21}	\bar{k}_{22}	\bar{k}_{211}	\bar{k}_{212}	\bar{k}_{222}
T_{10}	$k_{20} - ak_{21} + a^2 k_{211}$	$k_{21} - 2ak_{211}$	$k_{22} - ak_{212}$	k_{211}	k_{212}	k_{222}
T_{11}	k_{20}	$e^{-a} k_{21}$	k_{22}	$e^{-2a} k_{211}$	$e^{-2a} k_{212}$	k_{222}
T_{12}	k_{20}	k_{21}	$k_{22} - ak_{21}$	k_{211}	$k_{212} - 2ak_{211}$	$k_{222} - ak_{212} + a^2 k_{211}$
T_{111}	k_{20}	k_{21}	k_{22}	$k_{211} - ak_{21}$	k_{212}	k_{222}
T_{112}	k_{20}	k_{21}	k_{22}	k_{211}	$k_{212} - ak_{21}$	k_{222}
T_{122}	k_{20}	k_{21}	k_{22}	k_{211}	k_{212}	$k_{222} - ak_{21}$

Table 2.2(c)
Finite transformations of k

	\bar{k}_{10}	\bar{k}_{11}	\bar{k}_{12}	\bar{k}_{111}	\bar{k}_{112}	\bar{k}_{122}
T_{10}	$k_{10} - ak_{12} + a^2 k_{122}$	$k_{11} - ak_{112}$	$k_{12} - 2ak_{122}$	k_{111}	k_{112}	k_{122}
T_{11}	k_{10}	$k_{11} - ak_{12}$	k_{12}	$k_{111} - ak_{112} + a^2 k_{122}$	$k_{112} - 2ak_{122}$	k_{122}
T_{12}	k_{10}	k_{11}	$e^{-a} k_{12}$	k_{111}	$e^{-a} k_{112}$	$e^{-2a} k_{122}$
T_{211}	k_{10}	k_{11}	k_{12}	$k_{111} - ak_{12}$	k_{112}	k_{122}
T_{212}	k_{10}	k_{11}	k_{12}	k_{111}	$k_{112} - ak_{12}$	k_{122}
T_{222}	k_{10}	k_{11}	k_{12}	k_{111}	k_{112}	$k_{122} - ak_{12}$

Table 2.2(d)
Finite transformations of k

	\bar{k}_{20}	\bar{k}_{21}	\bar{k}_{22}	\bar{k}_{211}	\bar{k}_{212}	\bar{k}_{222}
T_{20}	$k_{20} - ak_{22} + a^2 k_{222}$	$k_{21} - ak_{212}$	$k_{22} - 2ak_{222}$	k_{211}	k_{212}	k_{222}
T_{21}	$k_{20} + ak_{10}$	$k_{21} - a^2 k_{12} + a(k_{11} - k_{22})$	$k_{22} + ak_{12}$	$k_{211} + a(k_{111} - k_{212}) + a^2(k_{222} - k_{112}) + a^3 k_{122}$	$k_{212} + a(k_{112} - 2k_{222}) - 2a^2 k_{122}$	$k_{222} + ak_{122}$
T_{22}	$e^a k_{20}$	$e^a k_{21}$	k_{22}	$e^a k_{211}$	k_{212}	$e^{-a} k_{222}$
T_{211}	k_{20}	$k_{21} + 2ak_{10}$	k_{22}	$k_{211} + a(2k_{11} - k_{22})$	$k_{212} + 2ak_{12}$	k_{222}
T_{212}	k_{20}	$k_{21} + ak_{20}$	$k_{22} + ak_{10}$	$k_{211} + ak_{21} + \frac{1}{2}a^2 k_{20}$	$k_{212} + ak_{11}$	$k_{222} + ak_{12}$
T_{222}	k_{20}	k_{21}	$k_{22} + 2ak_{20}$	k_{211}	$k_{212} + 2ak_{21}$	$k_{222} + ak_{22} + a^2 k_{20}$

formations $\exp(aV)$ becomes the time evolution operator if a is replaced by t . Equations (3.1) then simply restate the kinetic equations (2.1). (Of course, V is supposed known, while in the analysis just completed we have *determined* the U 's allowed for a given V .) Now the operator $\exp(tV)$ evolves an initial point into a trajectory in the space of x, k without changing the k 's. Taken together, all these trajectories constitute a flow because the coefficients r_i in (2.1) everywhere define a unique infinitesimal transformation $\exp(\delta t V)$. Each operator $\exp(aU)$ whose U satisfies the determining eqs. (2.8) will take a point P on such a trajectory and displace it in a transverse direction, by changing both x and k , giving an image point \bar{P} . If, with the same value of a , $\exp(aU)$ acts on another point P' of the original trajectory, it will carry this

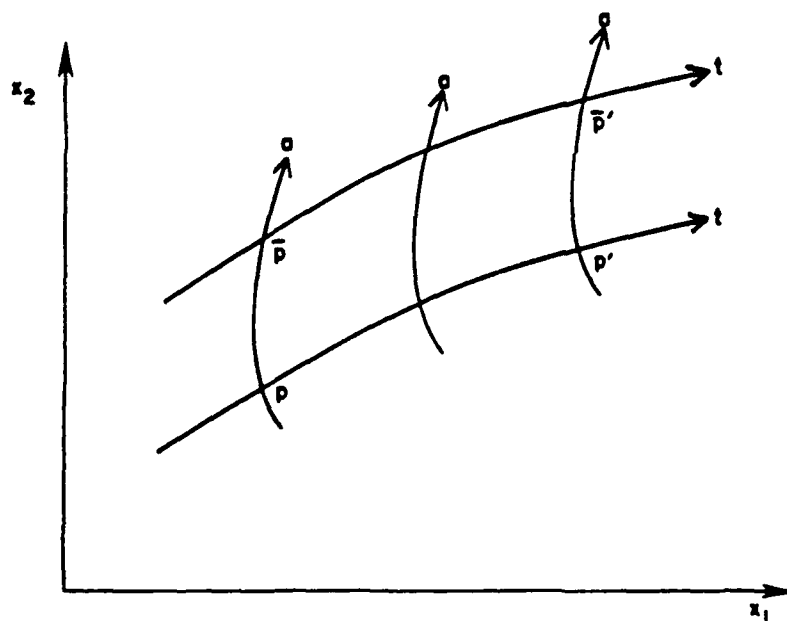


Fig. 2.1. Transformation flows $e^{aU}x$ transverse to evolution flows $e^{tV}x$. For each fixed value of the group parameter a , the transformation with generator U carries the evolving concentrations $x_i(t)$ into an altered set of evolving concentrations. The transformed concentrations obey a set of elementary kinetic equations with altered rate constants.

into an image point \bar{P}' . Because $\exp(\delta a U)$ everywhere defines a unique infinitesimal transformation and U is not proportional to V , the collection of all these trajectories produced by $\exp(aU)$ constitutes a flow transverse to the flow produced by the evolution operator. As indicated in fig. 2.1, the evolution operator will evolve the image point \bar{P} into a trajectory which will pass through \bar{P}' at the same time t that P is evolved into P' . The proof of this observation follows from the fact that in deriving (2.8) we have required that $\partial U / \partial t$ vanishes. Thus, (2.4) becomes

$$[U, V] = 0, \quad (3.3)$$

which implies that

$$\exp(tV) \exp(aU)(x, k) = \exp(aU) \exp(tV)(x, k). \quad (3.4)$$

When the generator U only approximately commutes with V , (3.4) will only hold approximately and the point obtained by transforming, then evolving, will not necessarily coincide with the point obtained by evolving, then transforming. This is the case for the generators $U_{ijj'}$, for example.

4. Conclusions

Inspection of eqs. (2.8) shows that if $U^{(-1)}$, $U^{(0)}$, and $U^{(1)}$ all vanish, then U does not act on the rate constants k . Thus, by determining all U with nonvanishing $U^{(-1)}$, $U^{(0)}$, $U^{(1)}$ whose $T(a)$ transform elementary rate equations into elementary rate equations, we have found *all* U generating one-parameter groups $\exp(aU)$ that transform elementary rate equations into elementary rate equations with different rate constants. The U_i and the U_{ij} have been determined exactly. In the $U_{ijj'}$, the functions governing the transformation of species concentrations have been determined to second order in the concentrations, and the functions governing the transformation of the rate constants have been exactly determined.

Throughout this and the following paper, two one-parameter groups are composed by allowing the second to act on the result obtained from the action of the first. Thus, if

$$x'_1 = \exp(bU_{112})x_1 = x_1 \exp(bx_2) \quad (4.1)$$

and

$$x'_2 = \exp(aU_{222})x_2 = x_2/(1 - ax_2), \quad (4.2)$$

then the effect of the transformation $\exp(bU_{112}) \exp(aU_{222})$ is to first shift the point with coordinates x_1, x_2 to the point with coordinates (x_1, x'_2) . It then moves this to the point with coordinates $(x'_1 = x_1 \exp(bx'_2), x'_2)$. Written as functions of the coordinates of the initial point, the coordinates of the final point are therefore

$$(x_1 \exp(b\{x_2/(1 - ax_2)\}), x_2/(1 - ax_2)). \quad (4.3)$$

From the one-parameter groups with operators $T_\alpha(a_\alpha) = \exp(a_\alpha U_\alpha)$, one may construct many-parameter groups $T_{\alpha\beta} \dots (a_\alpha, a_\beta, \dots) = \exp(a_\alpha U_\alpha) \exp(a_\beta U_\beta) \dots$ whenever

$$[U_\alpha, U_\beta] = \sum c_{\alpha\beta}^\nu U_\nu \quad (4.4)$$

for all α, β, ν . As all many-parameter groups may be obtained from one-parameter groups in this way, it may be concluded that our determination of the generators of all one-parameter groups that transform elementary rate equations into different elementary rate equations at once determines, exactly or approximately, all generators of many-parameter groups with this property. (In the following paper II, a list of such many-parameter groups is given for systems involving two chemical species.)

To conclude: In this paper, all generators of all one-parameter and all many-parameter groups of flows that transform elementary rate equations into elementary rate equations with different rate constants have been determined either exactly or approximately. A particularly simple generator basis has been chosen and the finite transformations obtained by exponentiating each generator have been determined.

Acknowledgements

The authors wish to thank Guang-Hui Xu and Gordon Ballentine for assistance with the calculations. This research was supported by the Air Force Office of Scientific Research.

References

- [1] H. Rabitz, *Chem. Rev.* 87(1987)101.
- [2] R. Cukoer, H. Levine and K. Schuler, *J. Comp. Phys.* 26(1978)1.
- [3] C. Box, W. Hunter and J. Hunter, *Statistics for Experimenters* (Wiley, New York, 1978).
- [4] C. Wulfman and H. Rabitz, *J. Phys. Chem.* 90(1986)2264.
- [5] L.M. Hubbard, C. Wulfman and H. Rabitz, *J. Phys. Chem.* 90(1986)2273.
- [6] Cf., for example, A. Cohen, *An Introduction to the Lie Theory of One-Parameter Groups* (Heath, Boston, 1911).

Appendix N

14. Global Sensitivity Analysis of Nonlinear Chemical Kinetic Equations Using Lie Groups; II. Some Chemical and Mathematical Properties of the Transformation Groups, C.E. Wulfman and H. Rabitz, J. Math. Chem., 3, 261 (1989).

**GLOBAL SENSITIVITY ANALYSIS OF NONLINEAR CHEMICAL
KINETIC EQUATIONS USING LIE GROUPS:
II. SOME CHEMICAL AND MATHEMATICAL PROPERTIES
OF THE TRANSFORMATION GROUPS**

C.E. WULFMAN

Department of Physics, The University of the Pacific, Stockton, CA 95207, USA

and

H. RABITZ

Department of Chemistry, Princeton University, Princeton, NJ 08544, USA

Received 14 December 1987
(in final form 2 January 1989)

Abstract

This paper establishes a number of properties of transformation groups that map elementary kinetic equations into new elementary kinetic equations with altered rate constants. The chemical significance of the transformations is assessed by applying them to systems involving two reacting species. There are then twelve one-parameter groups of mappings. Some mappings may be used to study the effects of changes in input/output fluxes on concentrations and their compensation by changes in other rate constants. A number of mappings transform nonlinear kinetics into approximately linear kinetics valid in regions larger than those obtained by standard methods. In some cases, the linearization is globally exact. Some mappings create lumped concentration variables and may be used to systematically reduce the number of manifest concentration variables in nonlinear, as well as linear, kinetic equations. The global mappings may be characterized by the functions of rate constants and functions of concentrations that they leave invariant. Although they produce large changes in rate constants and concentrations, none of these mappings change the topology of concentration phase plots as they map a phase plot determined by one set of initial conditions and rate constants into that determined by transformed initial conditions and rate constants. Metrical properties of the concentration maps generally depend upon the accuracy with which the group generators are approximated: systematic methods for their improvement are sketched.

1. Introduction

This paper is devoted to the assessment of key chemical and mathematical properties of the transformations determined in the preceding paper [1], hereafter referred to as I. To this end, we begin by considering kinetic systems with two constituents, present in concentrations x_1 and x_2 . Using the same notation for rate constants used in I, we will thus begin with transformations of the equations

$$\begin{aligned} dx_1/dt &= k_{10} + k_{11} x_1 + k_{12} x_2 + k_{111} x_1 x_1 + k_{112} x_1 x_2 + k_{122} x_2 x_2 \\ dx_2/dt &= k_{20} + k_{21} x_1 + k_{22} x_2 + k_{211} x_1 x_1 + k_{212} x_1 x_2 + k_{222} x_2 x_2. \end{aligned} \quad (1.1)$$

Section 2 applies a particular transformation of I to an exactly solvable pair of nonlinear kinetic equations with unstable solutions — a kinetic scheme used by Frank [2] as a model demonstrating the possibility of spontaneously developing optical activity in an initially achiral solution. Section 3 uses this same transformation to exactly linearize Frank's nonlinear rate equations and thereby leads to an indirect solution of them. Section 4 then considers a variety of transformations of these same rate equations and demonstrates that all the $T(a)$ of I act on Frank's equations to give transformed equations which possess unstable solutions.

Section 5 illustrates the application of the transformations of I to a kinetic system in which the linearizing transformation is not exact because the dependence of the group generator upon species concentrations has only been approximately determined. Unlike the usual methods of linearization which are accurate to $O(x^2)$, the linearization is accurate to $O(x^3)$. Section 6 is concerned with topological properties of the mappings in concentration space carried out by the transformations $T(a)$ of I. Two systems are defined to have qualitatively similar kinetics if their phase trajectories are topologically equivalent. It is shown that all the $T(a)$ of I convert phase curves into topologically equivalent phase curves. With this fact in hand, in section 7 it is shown how one may use the $T(a)$ to determine lumped concentration variables whose evolution is qualitatively similar to that of selected species of interest, yet governed by much simpler kinetic schemes. The $T(a)$ are also used to determine finite transformations of input/output fluxes that compensate for large changes in rate constants due to, for example, large temperature changes.

In section 8, the group generators established in I are used to determine functions of the rate constants that are left invariant by the transformations $T(a)$. This gives a global characterization of the mappings $x \rightarrow \bar{x} = T(a)x$, $k \rightarrow \bar{k} = T(a)k$, all of which make large changes in phase curves while leaving the topology of the phase curves unchanged. Section 9 determines the many-parameter groups whose transformations leave invariant the topology of the phase curves of a two-species system. Section 10 sets forth a method for improving the approximation to the transformed concentrations $\bar{x} = T(a)x$ one obtains when the generator U of $T(a)$ is approximate.

Section 11 sets forth an algorithm for improving the approximate generators used throughout the paper.

The final section, section 12, summarizes the results of this paper and I, and indicates directions for further investigation.

2. Solution of a set of nonlinear kinetic equations by transformation

To illustrate our transformation procedure, we use operators determined in I to change the value of the coefficients of the quadratic terms in the equations

$$dx_1/dt = px_1 + qx_1x_2 = k_{11}x_1 + k_{112}x_1x_2 \quad (2.1)$$

$$dx_2/dt = px_2 + qx_1x_2 = k_{22}x_2 + k_{212}x_1x_2.$$

Frank, and later Hochstim, used these equations with $p > 0$, $q < 0$ to model the chemical kinetics of a process in which an initially racemic mixture of two optical isomers with concentrations $x_1(t)$, $x_2(t)$ can spontaneously become optically active [2,3]. Although our purpose here is not a study of optical activity, reference to this interpretation will aid in understanding the transformations being used.

Perusing table 2.2 in I, we see that $T_{112}(\delta a)$ will change k_{212} to $k_{212} + \delta a k_{11}$, and that $T_{212}(\delta a)$ will change k_{112} to $k_{112} + \delta a k_{11}$. However, U_{112} and U_{212} do not commute; when a is finite, applying $T_{212}(a)$ to eqs. (2.1) after $T_{112}(a)$ gives a different result than applying $T_{112}(a)$ after $T_{212}(a)$. Neither sequence treats the two differential equations in the same manner. This leads us to use the generator $U = U_{112} + U_{212}$ in the operator $T(a) = \exp aU$ to change k_{112} and k_{212} . Using table 2.2 of I to evaluate the action of $\exp(\delta aU) = 1 + \delta a(U_{112} + U_{212})$ on x and k , we find that all k_{im} which vanish in (2.1) do not have their value changed, so we may drop many terms from $U_{112} + U_{212}$, specializing the generator to

$$U = x_1x_2\partial/\partial x_1 + x_1x_2\partial/\partial x_2 + k_{22}\partial/\partial k_{112} + k_{11}\partial/\partial k_{212}. \quad (2.2)$$

Evaluating $[V, U]$, one finds that this Lie generator exactly commutes with the evolution operator V for (2.1). If a is the group parameter in the transformation, one obtains for the transformed equations:

$$d\bar{x}_1/dt = p\bar{x}_1 + (q + ap)\bar{x}_1\bar{x}_2 \quad (2.3a)$$

$$d\bar{x}_2/dt = p\bar{x}_2 + (q + ap)\bar{x}_1\bar{x}_2.$$

In producing this result, we have considered the concentrations x_i to simply take on new values \bar{x}_i . On the other hand, we have explicitly indicated that $\bar{q} = q + ap$. This

highlights the effect of the transformation in changing the kinetic equation by changing rate constants. However, the explicit effect of the transformation on the species concentrations is also of importance. One finds by integrating equations (2.13) of I that $\exp(aU)$ acts on x to give, when $x_1 \neq x_2$:

$$\bar{x}_1 = \frac{x_1(x_1 - x_2)}{x_1 - x_2 \exp(aD)}, \quad (2.3b)$$

$$\bar{x}_2 = \frac{x_2(x_1 - x_2) \exp(aD)}{x_1 - x_2 \exp(aD)},$$

with

$$x_1 \neq x_2 \exp(aD), \quad (2.3c)$$

$$D = x_1 - x_2 = \bar{x}_1 - \bar{x}_2.$$

Note that for a given range of x_1 and x_2 , we have limited the range available to the parameter a so as to ensure that the finite transformation is 1:1 within the space of real x_1, x_2 , i.e. that $-\infty < x_1, x_2 < \infty$.

If $x_1 = x_2$, then one obtains

$$\bar{x}_1 = \frac{x_1}{1 - ax_1}, \quad \bar{x}_2 = \frac{x_2}{1 - ax_2}, \quad ax_1 \neq 1, \quad ax_2 \neq 1. \quad (2.3d)$$

It is not necessary to solve eqs. (2.3) above for the x to obtain the inverse transformation: because of the group property, the results will be the same as that obtained simply by changing a to $-a$ and interchanging the barred and unbarred variables. Thus, if $x_1 \neq x_2$:

$$x_1 = \frac{\bar{x}_1(\bar{x}_1 - \bar{x}_2)}{\bar{x}_1 - \bar{x}_2 \exp(-aD)}, \quad x_2 = \frac{\bar{x}_2(\bar{x}_1 - \bar{x}_2) \exp(-aD)}{\bar{x}_1 - \bar{x}_2 \exp(-aD)}. \quad (2.3e)$$

If $x_1 = x_2$, the inverse transformation is

$$x_1 = \frac{\bar{x}_1}{1 + a\bar{x}_1}, \quad x_2 = \frac{\bar{x}_2}{1 + a\bar{x}_2}. \quad (2.3f)$$

3. Linearization of the kinetics generating spontaneous optical activity

Returning to (2.3a), we note that if one sets $a = -q/p$ the coefficient of the quadratic terms in (2.3a) vanishes. This observation enables us to rather easily obtain

solutions of the original kinetic equations (2.1) in terms of elementary functions, for one may immediately integrate the linear equations obtained when the coefficient of the quadratic terms in (2.3a) vanishes. The result is

$$\bar{x}_1 = \bar{x}_1(0)\exp(pt), \quad \bar{x}_2 = \bar{x}_2(0)\exp(pt). \quad (3.1)$$

(Note that \bar{x}_1, \bar{x}_2 remain finite for all finite times so that the denominators in (2.3e) can only vanish as t approaches infinity.) Then, using the inverse transformations, one transforms the linearized equations back to the original nonlinear equations and thereby transforms (3.1) into their exact solution which, if $x_1(t_0) \neq x_2(t_0)$, is found to be

$$\begin{aligned} x_1(t) &= \frac{C_1(C_1 - C_2)\exp(pt)}{C_1 - C_2 \exp([q/p][C_1 - C_2]\exp[pt])} \\ x_2(t) &= \frac{C_2(C_1 - C_2)\exp(pt)\exp([q/p][C_1 - C_2]\exp[pt])}{C_1 - C_2 \exp([q/p][C_1 - C_2]\exp[pt])}, \end{aligned} \quad (3.2a)$$

where $C_i = \bar{x}_i(0)$. If $x_1(t_0) = x_2(t_0)$, then (2.3c) implies $C_1 = C_2 = C$, and the solutions of (2.1) are given by

$$x_1(t) = x_2(t) = \frac{C\exp(pt)}{1 - (q/p)C\exp(pt)}. \quad (3.2b)$$

These solutions agree with those obtained analytically by Frank using standard methods [2].

Note that the values of x_1 and x_2 at $t = 0$ are

$$\begin{aligned} x_1(0) &= \frac{C_1(C_1 - C_2)}{C_1 - C_2 \exp([q/p][C_1 - C_2])} \\ x_2(0) &= \frac{C_2(C_1 - C_2)\exp([q/p][C_1 - C_2])}{C_1 - C_2 \exp([q/p][C_1 - C_2])}, \end{aligned} \quad (3.2c)$$

when $x_1(0) \neq x_2(0)$. When the initial concentrations are equal, one has

$$x_1(0) = x_2(0) = \frac{C}{1 - (q/p)C}. \quad (3.2d)$$

Equations (2.1) have equilibrium (i.e. critical) points at $(0, 0)$ and $(-p/q, -p/q)$. As (x_1, x_2) approaches the unstable equilibrium point at $(-p/q, -p/q)$, the denomi-

nators in (3.2b) approach zero and x_1 and x_2 become infinite. Note, however, that it is impossible for any of the solutions to reach these equilibrium values from any other concentrations in any finite time.

It follows immediately from (2.3c) that if, when we start our clock ($t = 0$), the concentrations C_1 and C_2 of x_1 and x_2 are small but not identical, then

$$x_1(t) - x_2(t) = (C_1 - C_2) \exp(pt). \quad (3.3)$$

Thus, if any fluctuation in the concentrations of the D and L isomers leads to a momentary difference in these concentrations, this difference may grow exponentially with time. As Frank [2] first pointed out, because such fluctuations are to be expected on statistical grounds, a reaction system with kinetic equations (2.1), though started off with equal concentrations of D and L isomers, can lead to a preponderance of one isomer over the other. As will be seen in the following section, the methods we have developed enable one to systematically determine all other two-species elementary kinetic schemes which lead to the same result. However, we do not here provide methods for making a corresponding examination of systems where local concentration fluctuations and diffusion are involved. The interested reader is referred to the paper by Hochstim [3], who incorporated diffusion in the kinetics (2.1) and investigated the fluctuation dynamics of the system, as is necessary in any realistic theory of the spontaneous generation of optical activity by chemical means.

4. Distortions of kinetics generating spontaneous optical activity

The chemically significant feature of the kinetics in the previous two sections is the instability of solutions in which the concentrations of D and L isomers are equal: if these concentrations momentarily become unequal at time t_0 , then thereafter

$$x_1(t) - x_2(t) = \{x_1(t_0) - x_2(t_0)\} \exp(t - t_0)p. \quad (4.1)$$

It is instructive to see what the invariance transformations do to the kinetic equations (2.1) and to the time evolution of this difference. To avoid confusion with the transformation of the previous section, we shall in this section write

$$\tilde{k} = T(a)k, \quad k = T(-a)\tilde{k}, \quad \tilde{x} = T(a)x, \quad x = T(-a)\tilde{x}. \quad (4.2)$$

We first consider the exact invariance transformations T_{10} , T_{11} , T_{12} . Letting $x = T_{10}(-a)\tilde{x} = (x_1 - a, x_2)$, we find using table 2.2 of I that

$$\tilde{k}_{10} = -ak_{11}, \quad \tilde{k}_{12} = -ak_{112}, \quad \tilde{k}_{22} = k_{22} - ak_{212}, \quad (4.3a)$$

while all other k 's are unchanged. Also,

$$\tilde{x}_1 - \tilde{x}_2 = (C_1 - C_2) \exp(pt) + a. \quad (4.3b)$$

Thus, $T_{10}(-a)$ converts the Frank equations into

$$\dot{\tilde{x}}_1 = -ap + p\tilde{x}_1 - aq\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2, \quad \dot{\tilde{x}}_2 = (p - aq)\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2. \quad (4.4)$$

It is evident from (4.3b) that these new equations also possess unstable solutions in the same sense as do eqs. (3.1).

Next, let $(x, k) = T_{11}(-a)(\tilde{x}, \tilde{k})$. Using table 2.2 of I, one finds

$$\tilde{x}_1 = x_1 e^a, \quad \tilde{x}_2 = x_2 \quad (4.5)$$

and

$$\dot{\tilde{x}}_1 = p\tilde{x}_1 + q\tilde{x}_1\tilde{x}_2, \quad \dot{\tilde{x}}_2 = p\tilde{x}_2 + e^{-aq}\tilde{x}_1\tilde{x}_2.$$

Thus, for these equations one has

$$\tilde{x}_1 - \tilde{x}_2 = (C_1 \exp(a) - C_2) \exp(pt). \quad (4.6)$$

Applying $T_{12}(-a)$, one obtains

$$\tilde{x}_1 - \tilde{x}_2 = x_1 - x_2 - ax_2. \quad (4.7)$$

which grows exponentially as t becomes large. The transformed kinetic equations are

$$\dot{\tilde{x}}_1 = p\tilde{x}_1 + q(1 + a)\tilde{x}_1\tilde{x}_2 - q(a + a^2)\tilde{x}_2^2, \quad \dot{\tilde{x}}_2 = p\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2 - aq\tilde{x}_2^2. \quad (4.8)$$

We turn next to the action of transformations that only leave the kinetic equations approximately invariant.

Using table 2.2 of I to determine the action of $T_{111}(-a)$, one finds:

$$\tilde{x}_1 = \frac{x_1}{1 - ax_1}, \quad \tilde{x}_2 = \tilde{x}_2 \quad (4.9)$$

$$\tilde{x}_1 - \tilde{x}_2 = \left\{ \frac{C_1}{1 - aC_1 \exp(pt)} - C_2 \right\} \exp(pt)$$

and that

$$\tilde{k}_{111} = ak_{11} = ap. \quad (4.10)$$

The corresponding differential equations are

$$\begin{aligned}\dot{\tilde{x}}_1 &= p\tilde{x}_1 + ap\tilde{x}_1^2 + q\tilde{x}_1\tilde{x}_2 + O(x^3) \\ \dot{\tilde{x}}_2 &= p\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2 + O(x^3).\end{aligned}\quad (4.11)$$

$T_{112}(-a)$ gives

$$\tilde{x}_1 = x_1 e^{ax_2}, \quad \tilde{x}_2 = x_2, \quad (4.12)$$

$$\tilde{x} - \tilde{x}_2 = \{C_1 \exp(aC_2 \exp(pt)) - C_2\} \exp(pt)$$

and

$$\tilde{k}_{112} = k_{112} + ak_{22} = q + ap. \quad (4.13)$$

The transformed differential equations are

$$\begin{aligned}\dot{\tilde{x}}_1 &= p\tilde{x}_1 + (q + ap)\tilde{x}_1\tilde{x}_2 + O(x^3) \\ \dot{\tilde{x}}_2 &= p\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2 + O(x^3)\end{aligned}\quad (4.14)$$

Finally, $T_{122}(-a)$ yields the transformed solutions

$$\tilde{x}_1 = x_1 - ax_2^2, \quad \tilde{x}_2 = x_2, \quad \tilde{k}_{122} = a(2k_{22} - k_{11}) = ap, \quad (4.15)$$

so that

$$\tilde{x}_1 - \tilde{x}_2 = (C_1 - C_2)\exp(pt) + aC_2^2 \exp(2pt) \quad (4.16)$$

depicts the time evolution of concentration differences for the resulting solutions of the equation

$$\begin{aligned}\dot{\tilde{x}}_1 &= p\tilde{x}_1 + q\tilde{x}_1\tilde{x}_2 + ap\tilde{x}_2^2 + O(x^3) \\ \dot{\tilde{x}}_2 &= p\tilde{x}_2 + q\tilde{x}_1\tilde{x}_2 + O(x^3).\end{aligned}\quad (4.17)$$

It will be noted that although these various transformations lead to equations with little self-evident relationship to the Frank equations, all the solutions have the property that they develop exponential growth of the difference between concentrations. By acting successively with the twelve different transformations of table 2.2 of I, one obtains from the Frank equations a twelve-parameter family of kinetic

equations, all of which possess similarly unstable solutions. In sections 6 and 7, we will establish the exact sense in which this property of our transformations is a general one.

5. Transformations of Lotka–Volterra systems

The example of sections 2 and 3 is somewhat misleading because the kinetic system possesses no separatrix in the phase plane and because we were able to use an exact invariance transformation to linearize the rate equations. In this section, we investigate the more typical example provided by the rate equations of Lotka and Volterra [4,5]. They can always be reduced to the special case [6]

$$\dot{x}_1 = p(x_1 - x_1 x_2), \quad \dot{x}_2 = -q(x_2 - x_1 x_2), \quad (5.1)$$

which has critical points at (0,0) and (1,1). If one rewrites these about the second singular point by making the substitution

$$x_1 = y_1 + 1, \quad x_2 = y_2 + 1,$$

then they become

$$\dot{y}_1 = p(-y_2 - y_1 y_2), \quad \dot{y}_2 = -q(-y_1 - y_1 y_2). \quad (5.2)$$

In this section we will, for simplicity, consider $p = q = 1$.

As we wish to allow the k_{im} to vary, we consider (5.1) to be a special case of the equations

$$\dot{x}_1 = k_{11} x_1 + k_{112} x_1 x_2, \quad \dot{x}_2 = k_{22} x_2 + k_{212} x_1 x_2, \quad (5.3)$$

with $k_{11} = 1, k_{112} = -1, k_{22} = -1, k_{212} = 1$.

Similarly, (5.2) is a special case of the equations

$$\dot{y}_1 = k_{12} y_2 + k_{112} y_1 y_2, \quad \dot{y}_2 = k_{21} y_1 + k_{212} y_1 y_2, \quad (5.4)$$

with $k_{12} = -1 = k_{112}, k_{21} = 1 = k_{212}$.

Comparing eqs. (5.3) with eqs. (2.1), we find that the generator U of (2.2) is the generator of a transformation that will linearize (5.3). However, in this case the equations are only approximately invariant under the transformation: Evaluating $[V, U]$, one obtains as the remainder a $W^{(2)}$ term with components

$$(w_1, w_2) = (-2x_1^2 x_2, 2x_1 x_2^2). \quad (5.5)$$

This remainder is of higher order in x than that obtained in the standard local linearization which simply neglects terms of $O(x^2)$.

We shall henceforth use the term *regional* to denote an approximation, such as this linearization, whose error terms are of order x^3 or greater.

Equations (5.4) may be linearized in a manner similar to that used for eqs. (5.3). Using table 2.2 of I, one finds that to linearize (5.4) it is necessary to make use of all the generators quadratic in x . Utilizing the infinitesimal transformations as before, one finds that a transformation with generator

$$U_{111} + U_{122} - U_{122} - U_{211} + U_{212} + U_{222} \quad (5.6)$$

will have the desired effect. Because many of the k 's that are zero do not have their values altered by the T_{ijk} , the generator (5.6) may be simplified to

$$U = (y_1^2 + y_1 y_2 - y_2^2) \partial / \partial y_1 + (y_2^2 + y_1 y_2 - y_1^2) \partial / \partial y_2 \\ + (k_{12} - 2k_{21}) \partial / \partial k_{112} + (k_{21} - 2k_{12}) \partial / \partial k_{212} . \quad (5.7)$$

Evaluating $[V, U]$, one finds that in the remainder

$$w_1 = k_{112} y_1^3 + k_{212} y_1^2 y_2 - (k_{112} + 2k_{212}) y_1 y_2^2 + k_{112} y_2^3 \\ w_2 = k_{212} y_1^3 - (2k_{112} + k_{212}) y_1^2 y_2 + k_{112} y_1 y_2^2 + k_{212} y_2^3 . \quad (5.8)$$

Acting with $\exp(aU)$ on the equations, they are, respectively, converted into

$$\dot{\bar{x}}_1 = k_{11} x_1 + (k_{112} + ak_{22}) \bar{x}_1 \bar{x}_2 + O(x^3) \\ \dot{\bar{x}}_2 = k_{22} x_2 + (k_{212} + ak_{11}) \bar{x}_1 \bar{x}_2 + O(x^3) \quad (5.9)$$

and

$$\dot{\bar{y}}_1 = k_{12} \bar{y}_2 + (k_{112} - 3ak_{21}) \bar{y}_1 \bar{y}_2 + O(y^3) \\ \dot{\bar{y}}_2 = k_{21} \bar{y}_1 + (k_{212} - 3ak_{12}) \bar{y}_1 \bar{y}_2 + O(y^3) . \quad (5.10)$$

The effect of the error terms will be discussed in sections 10 and 11.

Setting $a = -1$ in (5.9) and $a = -1/3$ in (5.10), one obtains linear equations whose solutions are, respectively,

$$\bar{x}_1 = C_1 \exp(t), \quad \bar{x}_2 = C_2 \exp(-t) \quad (5.11)$$

and

$$\begin{aligned}\bar{y}_1 &= C_1 \cos(t) - C_2 \sin(t) \\ \bar{y}_2 &= C_2 \cos(t) + C_1 \sin(t).\end{aligned}\tag{5.12}$$

Using the inverse transformations developed in section 2, one converts (5.11) into approximate solutions of the Lotka-Volterra equations. One finds, as before,

$$\begin{aligned}x_1(t) &= \frac{\bar{x}_1(\bar{x}_1 - \bar{x}_2)}{\bar{x}_1 - \bar{x}_2 \exp\{-a(\bar{x}_1 - \bar{x}_2)\}}, \\ \text{or if } x_1 &= x_2, \quad \frac{\bar{x}_1}{1 + a\bar{x}_1} \\ x_2(t) &= \frac{\bar{x}_2(\bar{x}_1 - \bar{x}_2) \exp\{-a(\bar{x}_1 - \bar{x}_2)\}}{\bar{x}_1 - \bar{x}_2 \exp\{-a(\bar{x}_1 - \bar{x}_2)\}}, \\ \text{or if } x_1 &= x_2, \quad \frac{\bar{x}_2}{1 + a\bar{x}_2},\end{aligned}\tag{5.13}$$

where \bar{x}_1 and \bar{x}_2 are given by (5.11) in the vicinity of the origin. The range of a must be restricted to ensure that the transformations are 1:1 on the reals.

In the vicinity of (1,1), one uses the transformation with generator U given in (5.7) to obtain the transformed variables. We may take advantage of the fact that the commutator of any two of the generators composing this U either vanishes or is of order y^3 . As a result, to order y^3 we may write $\exp(aU)$ as a product $\exp(aU_{111})\exp(aU_{112}) \dots \exp(aU_{222})$. Proceeding in such a manner, we find

$$\begin{aligned}y_1 &= \frac{\bar{y}_1 + a\bar{y}_2^2}{1 + a(\bar{y}_1 + \bar{y}_2) + a^2(\bar{y}_1^2 + \bar{y}_2^2)} \\ y_2 &= \frac{\bar{y}_2 + a\bar{y}_1^2}{1 + a(\bar{y}_1 + \bar{y}_2) + a^2(\bar{y}_1^2 + \bar{y}_2^2)},\end{aligned}\tag{5.14}$$

with \bar{y}_1, \bar{y}_2 given as functions of t by (5.12). Note that y_1 and y_2 are single valued functions of \bar{y}_1, \bar{y}_2 and a for the allowed range of these variables. Hence, as \bar{y}_1 and \bar{y}_2 are cyclic functions of t , y_1 and y_2 must be cyclic in t . This has the consequence that the closed curves which are the phase plane plots of \bar{y}_1, \bar{y}_2 are mapped into closed phase curves of y_1, y_2 .

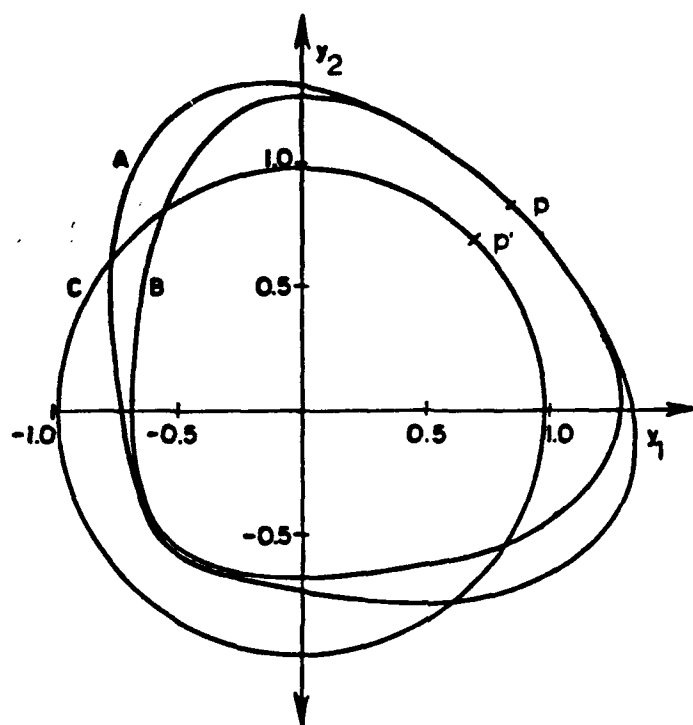


Fig. 5.1. Global approximation to a phase trajectory of the Lotka–Volterra equation. The trajectory B of the Lotka–Volterra equation is approximated by the trajectory A defined by (5.14). C is the reference circle defined by (5.12).

Because the U of (5.7) is only approximate, eqs. (5.14) do not yield exact solutions of the rate equations when a is assigned the prescribed value of $-1/3$. In fig. 5.1, an approximate phase trajectory (A) determined by (5.14) and (5.12) is compared with the trajectory (B) obtained by numerical integration of the Lotka–Volterra equations. The corresponding trajectory of the linearized equations is plotted in the figure as (C). In obtaining these trajectories, the initial point p was used to determine p' on the reference circle defined by (5.12). In section 10, a method is developed for improving the approximate trajectory in the region of any point of interest.

6. Transformation of phase trajectories: Topological invariants

A key feature of any kinetic system is the behaviour of its phase portrait [6–8]. (We shall use the term phase portrait when we are referring to trajectories in the vicinity of singular points in the phase space $\{x\}$.) As a result, it is important to

investigate the way in which these portraits are affected by the transformations we have obtained. To introduce this study, we carry out a standard investigation of the phase portraits of (1.1). When $q \neq 0$, the right-hand sides of the equations vanish for $x_1 = x_2 = 0$, and for $x_1 = x_{10} = -p/q$, $x_2 = x_{20} = -p/q$. Only the first critical point persists if $q = 0$. In the region of the critical point at the origin, the solutions of the equations are

$$x_1(t) = x_1(0)\exp(pt), \quad x_2(t) = x_2(0)\exp(pt) \quad (6.1)$$

and the phase portrait consists of trajectories fleeing the origin, an improper node. (Of course, on interpreting x_1 and x_2 as species concentrations, one sees that the trajectories on which either of these variables become negative have no direct chemical relevance.) The invariance transformations of section 2 merely distort these trajectories as they recede from the origin, but none of the transformations changes the topological classification of the portrait.

We next turn to an investigation of the phase portraits in the region of the second critical point at $(-p/q, -p/q)$. Letting $y = x - (-p/q, -p/q)$ and expressing the equations about this second critical point yields

$$dy_1/dt = -py_2 + qy_1y_2, \quad dy_2/dt = -py_1 + qy_1y_2. \quad (6.2)$$

The secular equation of the linear part of this system is

$$\text{Det} \begin{pmatrix} -\lambda & -p \\ -p & -\lambda \end{pmatrix} = 0 = \lambda^2 - p^2. \quad (6.3)$$

It will be noted that the roots are independent of q . Since these roots determine the phase portrait, it is evident that the portrait is independent of q whenever y is well defined, i.e. for $q \neq 0$. The portrait is that of a saddle point. Applying the transformations of table 2.2 of I to y_1 and y_2 , one finds, as in the previous case, that the topological classification of the portrait is unchanged.

The Lotka-Volterra system of section 5 has an unstable saddle point at the origin, and a stable center at $(1, 1)$. Thus, the portrait in the region of the first critical point and that in the region of the second critical point are of radically different topological type. (Although only the latter is of direct chemical interest, we shall for illustrative purposes consider them both.) Applying the transformations of table 2.2 of I to the variables x in eqs. (5.1) and the variables y in eqs. (5.2), one finds that neither phase portrait may be changed into the other or into a portrait of a different topological classification.

It is a difficult task to determine all possible phase portraits for just two elementary kinetic equations. One must first locate all stationary points $dx_1/dt = 0 = dx_2/dt$.

This is equivalent to investigating and classifying all possible intersections of the pair of conics defined by setting the right-hand sides of (1.1) to zero, which if they are not identical, may intersect at 4, 3, 2, 1 or no points. To then investigate the action of all the transformations in table 2.2 of I on each phase portrait is a task one would like to avoid. In the following paragraphs, we determine the effects of the transformations on the topological properties of all possible phase portraits without proceeding on a case by case basis, and without confining the system to a phase space of two dimensions.

In the examples of this and previous sections, we have seen transformations of kinetic equations that have preserved qualitative features of the solutions of the equations even though they may have greatly changed the concentrations and rate constants, and hence the equations themselves. All transformations of the equations introduced by Frank were found to preserve the instability of the solutions with equal concentrations of D and L isomers portrayed in the phase portrait of the untransformed system. All transformations of the cyclic solutions of the Lotka-Volterra equations in the region of their critical point gave rise to cyclic solutions, and all transformations of the non-cyclic solutions in the region of their critical point yielded non-cyclic solutions. None of the transformations in the examples altered the topological classification of a critical point.

Let us therefore address the question of whether it is true in general that our invariance transformations change phase trajectories in such a manner as to preserve the topological properties of the trajectories everywhere in the phase space.

First of all we ask whether the operators $\exp(aU)$ always transform closed phase curves into closed phase curves, and open phase curves into open phase curves? The answer to this question is yes, for the following reasons. The polynomial form of the coefficient functions in the generators U ensures that the coefficients $h_i(x)$ are single valued differentiable, indeed analytic, functions, and this is true even when the polynomials are only approximations to the exact h_i . Now, at each point in the phase space the infinitesimal shift in x, k brought about by an infinitesimal transformation with parameter δa is given by $\delta a Ux, \delta a Uk$. Thus, at each point in phase space ($-\infty < x_i < \infty$ for all i), our infinitesimal transformations define a unique shift of the point, that is to say, they are local diffeomorphisms. We have not allowed finite transformations that shift x_i outside this same range. Since the finite transformations $T(a)$ are compounded of a succession of infinitesimal transformations $T(\delta a)$ such that $a = \int \delta a$, for each value of a they also determine unique motions of each point in x, k, t space as long as x, k, t remain real. Thus, first of all, for all a within the allowed range, the transformations carried out by the operators $\exp(aU)$, in addition to being unique and having a unique inverse, vary smoothly from point to point and carry contiguous regions in x, k, t space into contiguous regions, and discontinuous regions into discontinuous regions — that is to say, they are local diffeomorphisms of the space of x, k, t [7]. Second, because we do not allow values

of the group parameter which would transform any variable outside the reals, the transformations are global diffeomorphisms of the space of real x, k, t . In addition, the transformations are time independent so that they are diffeomorphisms of x, k space. Finally, the transformations are such that as x varies with t , k does not vary. It follows from this that as t progresses and a phase trajectory and its transformed image develop (a being held fixed), if it should happen that the phase point returns to its initial position, then its transformed image will also return to its corresponding initial position. Thus, a closed phase curve is mapped into a closed phase curve. In a similar way, one argues that because the transformations are t -independent diffeomorphisms of x, k, t space, they carry discontinuous regions of phase space into discontinuous regions, and hence transform open phase curves into open phase curves.

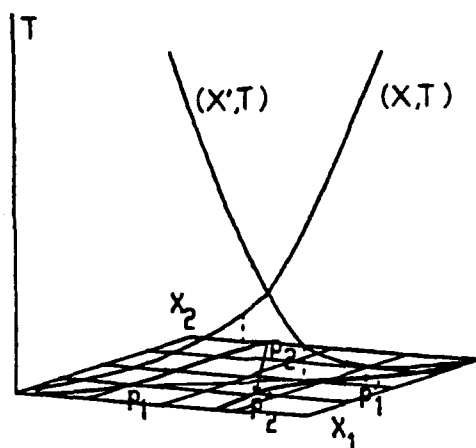
It is evident from this discussion that our transformations allow us to determine changes in rate constants that will leave an initially oscillatory reaction oscillatory and an initially non-oscillatory reaction non-oscillatory. Any transformation compounded of transformations $\exp(aU')$, each of whose generators are of the form

$$U' = \sum c_m(k) U_m, \quad (6.4)$$

will have this property when acting on the x_i if the U_m are those determined in section 2, and the c_m are smooth functions of k .

In the usual topological classification of phase portraits and phase curves, the direction of motion as t increases is also a topological invariant. Hence, we next investigate whether any of our changes in rate constants invert the direction of motion along a phase curve.

Inspecting table 2.2 of I, one finds that none of its transformations can have such an effect. The underlying reason for this is perhaps most clearly seen with the aid of fig. 6.1, which purports to depict a solution curve in x_1, x_2, t space and its projections onto x_1, x_2 phase space, together with another curve in this phase space. Suppose that at times t_1 and t_2 the points P_1 and P_2 are marked on a trajectory of growing concentrations. Suppose that for a given value a' of the group parameter it were to happen that $\exp(aU)$ were to map P_1 into P'_1 — and that for the same value of the group parameter, P_2 is carried into P'_2 , a point where \bar{x}_1 and \bar{x}_2 have smaller values than at P'_1 . The arrows are drawn in to indicate how, as one increases the group parameter from 0 to a' , the transformed points move away from the original trajectory. It will be noted that these lines cross at some intermediate value of a . However, if this were to happen, then for larger values of a the transformation would have to carry the point of crossing into both P'_1 and P'_2 — and the inverse transformation would have to carry the point to both P_1 and P_2 . Because our generators U have single valued functions for their coefficients, the infinitesimal transformations are everywhere unique and all this is impossible. In short, it is impossible to convert the first phase trajectory into the second using any of our $T(a)$.



An Impossible Mapping

Fig. 6.1. An impossible mapping. Two curves $x(t)$ can not be mapped into one another by any of the transformations considered in this paper if one depicts concentrations that increase with time and the other depicts concentrations that decrease with time at the same time.

The argument just given evidently fails if the phase space is more than two dimensional, for then the lines $P_1 P_1'$ and $P_2 P_2'$ need not intersect. In such cases, we may consider an initial phase trajectory which develops in one direction as t increases, and a nearby phase trajectory obtained from the first by a transformation with operator $T(a)$ – a trajectory which by hypothesis evolves in the opposite direction. If two such curves exist we can, from arguments of continuity in the group parameter a , conclude that between them lie two similar curves that are connected by an infinitesimal transformation $T(\delta a)$ and that between these two curves lies a curve along which points do not move with t . Thus, along this intermediate curve all \dot{x}_i vanish. We now prove that in the region of this intermediate curve, T cannot change any of the rates \dot{x}_i . The effect upon x_i of the infinitesimal transformation with generator U is to convert x_i to $\bar{x}_i = x_i + \delta a h_i(x)$. This induces a transformation of dx_i/dt to

$$d\bar{x}_i/dt = \frac{d}{dt}(x_i + \delta a h_i(x)) = \dot{x}_i + \delta a \sum \dot{x}_j \partial h_i / \partial x_j. \quad (6.5)$$

As \dot{x}_i and all the other \dot{x}_j vanish on the intermediate curve, we see that in its infinitesimal neighbourhood T cannot change any of the \dot{x} 's and so cannot change the direction of motion along any trajectory. It follows from continuity in the group parameter a that T is unable to transform any trajectory into a trajectory developing oppositely in time.

The observations so far made in this section may be subsumed in the general observation that because our transformations are, for each allowed value of the group parameter a , diffeomorphisms of the space of x, k that keep dk/dt zero, they transform phase trajectories into topologically equivalent phase trajectories [7].

It is important to note that because even our approximate invariance transformations are local and global diffeomorphisms, all the above statements hold true even for them. Of course, when one uses approximate invariance transformations, one converts exact solutions into approximate solutions and hence, usually, converts exact phase trajectories of one kinetic system into approximate phase trajectories of another. Nevertheless, increasing the accuracy of the approximation by increasing the number of terms in the power series approximation to the $h_i(x)$ will not alter the topology of the target curve, which is completely determined by the topology of the untransformed solution curve. *Thus, for all the transformations we allow, the evolution of the original system and the evolution of the transformed systems are qualitatively similar in a well-defined sense: their phase curves are topologically indistinguishable.* The topology of the phase curves is, in the standard sense which includes the direction of motion, an invariant of our transformations.

To sum up our observations to this point: the methodology and conceptions we have described enable one to establish well-defined qualitative relations, as well as quantitative relations, between the behaviour of kinetic systems with different rate constants. Because one may transform many rate constants to zero, the conceptions are also applicable to studies relating the global behaviour of systems with complex kinetics to the behaviour of systems with simpler kinetics – and vice versa.

7. Lumping and flux control

Both in the analysis and in the utilization of kinetic studies of complex reacting systems, one often tries to simplify the kinetic scheme by 'lumping' a number of reactions into one, thus submerging a part of the detailed elementary kinetics. For this goal, it is necessary that the reactions retained in the kinetic scheme proceed at least qualitatively, as they would if the submerged reactions were taken into account. Because we are assured that our transformations do not change the qualitative behaviour of a kinetic system, it is worthwhile to determine whether they can be used to determine lumpings. Sometimes a lumping is only possible because the initial concentrations satisfy some special relationship, and sometimes it is only possible because some kinetic coefficients are confined to some special range of values. While the methods developed in this article can be of help in studying both these situations, here we wish only to deal with the use of the methods in the global analysis of kinetic systems. That is to say, we are here concerned only with the consequences of large changes in kinetic coefficients and with consequences that are independent of initial concentrations.

To exemplify our approach to lumping, we begin by considering the inverse process, that of sophisticating one member of a set of rate equations — an equation that happens to involve only one species. Consider the general elementary kinetic scheme involving only one species:

$$\dot{x}_1 = g_{10} + g_{11} x_1 + g_{111} x_1^2. \quad (7.1)$$

We may suppose that while this reaction is proceeding, another reaction involving x_2 is also proceeding independently. Now the concentration x_1 necessarily evolves in a non-oscillatory manner. Acting on (7.1) with any of the twelve transformations $T(a)$ of table 2.2 of I will give a one-parameter family of two-component kinetic systems in which \bar{x}_1 's evolution is also non-oscillatory. Acting with each of the twelve transformations in succession will give a twelve-parameter family of such kinetic systems.

The lumped variable \bar{x}_1 resulting from these transformations will in general be a complicated function of x_1 and the other concentrations, but as the group parameters become smaller and smaller, it will come closer and closer to being x_1 . Even though \bar{x}_1 makes large excursions and the kinetic coefficients may be greatly altered, the evolution of \bar{x}_1 for all members of this twelve-parameter family of reactions is globally, i.e. topologically, equivalent to that of the lumped system (7.1). All this is to say that \bar{x}_1 will behave qualitatively as though it were x_1 .

Consider now the process involved in eliminating a concentration variable from a kinetic equation using transformations $x_1, x_2 \rightarrow \bar{x}_1, \bar{x}_2$. It might appear at first sight that with a twelve-parameter family of lumping transformations available, one could lump away just about any variable in a reaction without changing the topology of the phase trajectories. In this connection, an example involving the lumping of three species into two may be revealing. Consider the reactions



and suppose that A is being supplied at rate k_0 while C is being supplied at rate k_3 . Let us try to transform away the intermediate species in the final reaction. Assigning the index i antilexicographically, the associated kinetic equations are

$$\begin{aligned} \dot{x}_3 &= k_0 + 2k_{-1} x_2 - k_1 x_3 x_3 = k_{30} + k_{32} x_2 + k_{333} x_3 x_3 \\ \dot{x}_2 &= -k_{-1} x_2 - 2k_{-2} x_1 + k_1 x_3 x_3 - 2k_2 x_2 x_2 \\ &= k_{22} x_2 + k_{21} x_1 + k_{233} x_3 x_3 + k_{222} x_2 x_2 \\ \dot{x}_1 &= k_3 - k_{-2} x_1 + k_2 x_2 x_2 = k_{10} + k_{11} x_1 + k_{122} x_2 x_2. \end{aligned} \quad (7.3)$$

We wish to carry out a transformation $x \rightarrow \bar{x}$, $k \rightarrow \bar{k}$ which will eliminate \bar{x}_2 from the last reaction. Perusing table 2.1 of I and taking into account the fact that a number of k 's vanish in the intermediate and final reactions, we see that U_{12} is the only generator available for this purpose. Table 2.2 then indicates that

$$\bar{k}_{122} = k_{122} + ak_{222}, \quad \bar{k}_{12} = a(k_{22} - k_{11}) - a^2k_{21}. \quad (7.4)$$

Thus, on setting $a = -k_{122}/k_{222} = k_2/-k_2 = -1$ we can transform k_{122} to zero — but we will also, in general, create a nonzero k_{12} . Again perusing table 2.1 of I, we find that we can not find another transformation that will eliminate the unwanted \bar{k}_{12} . It follows that we can only attain our desired end if it should happen that the value of a which makes \bar{k}_{122} vanish also makes \bar{k}_{12} vanish. This will happen only if

$$(k_{122}/k_{222})k_{21} + k_{22} - k_{11} = -(k_{-2} + k_{-1}) = 0. \quad (7.5)$$

As untransformed rate constants can not be negative, it is evident (7.5) can only be satisfied if we can replace the k 's by some negative \bar{k} 's by means of some further transformation. Perusing tables 2.1 and 2.2 of I, one finds that a candidate for such a transformation is provided by $T_{111}(b)$. It acts on k_{11} to give $k_{11} + 2bk_{10}$ so that the term in (7.5) which must vanish becomes $-(k_{-2} + k_{-1} + 2bk_3)$. Thus, by setting $b = -(k_{-1} + k_{-2})/2k_3$, the lumping becomes possible. The only other effect of $T_{111}(b)$ on the final reaction is to convert x_1 to $x_1/(1 + bx_1)$. Applying T_{111} after U_{12} , the lumped concentration variable will be $\bar{x}_1 = (x_1 + ax_2)/(1 + b\{x_1 + ax_2\})$. The other concentrations x_2 and x_3 are unaffected. The kinetics of the final reaction will become

$$\begin{aligned} \dot{\bar{x}}_1 &= k_{10} + \bar{k}_{11}\bar{x}_1 + \bar{k}_{111}\bar{x}_1\bar{x}_1 \\ \bar{k}_{11} &= k_{11} + 2bk_{10} + ak_{21}, \quad \bar{k}_{111} = ak_{11} + a^2k_{12} + bk_{21}. \end{aligned} \quad (7.6)$$

Lumped concentration variables are also of use in another setting, in which one wishes the lumped variables to behave qualitatively like the original concentrations. It is a common experience that heat produced in the course of a chemical reaction may affect reaction rates (and, as a result, product composition) by changing unimolecular rate constants k_{ij} and bimolecular rate constants k_{ijk} . One commonly controls such reactions by adjusting cooling rates and by adjusting concentrations and rates of supply of reagents. For reactions involving two species, the extent to which time-independent reaction fluxes and concentration changes may be so used can be determined with the aid of table 2.1 of I. Perusing the table, one sees that only the generators U_{10} and U_{11} have nonzero values of g_{10} and g_{20} . Thus, only transformations using them can adjust the fluxes k_{10} and k_{20} . The most general allowed generator available for such purposes is a linear combination of these six generators of the form

$$U = \sum c_m(k) U_m. \quad (7.7)$$

Using table 2.1 of I one finds that an infinitesimal transformation with this generator has the following effect on the flux k_{10} and the rate constants $k_{1\mu}$:

$$\begin{aligned} \delta k_{10} &= \delta a \{-c_{10} k_{11} + c_{11} k_{10} + c_{12} k_{20} - c_{20} k_{12}\} \\ \delta k_{11} &= \delta a \{-2c_{10} k_{111} + c_{12} k_{21} - c_{20} k_{112} - c_{21} k_{12}\} \\ \delta k_{12} &= \delta a \{-c_{10} k_{112} + c_{11} k_{12} + c_{12} (k_{22} - k_{11}) - 2c_{20} k_{122} - c_{22} k_{12}\} \\ \delta k_{111} &= \delta a \{-c_{11} k_{111} + c_{12} k_{211} - c_{21} k_{121}\} \\ \delta k_{112} &= \delta a \{c_{12} (k_{212} - 2k_{111}) - 2c_{21} k_{112} - c_{22} k_{112}\} \\ \delta k_{122} &= \delta a \{c_{11} k_{122} + c_{12} (k_{222} - k_{112}) - 2c_{22} k_{122}\}. \end{aligned} \quad (7.8)$$

The associated changes in concentrations are

$$\begin{aligned} \delta x_1 &= \delta a \{c_{10} + c_{11} x_1 + c_{12} x_2\} \\ \delta x_2 &= \delta a \{c_{20} + c_{21} x_1 + c_{22} x_2\}. \end{aligned} \quad (7.9)$$

A similar set of relations can be written for the flux k_{20} and rate constants $k_{2\mu}$. To negate the effects of infinitesimal temperature-driven changes in the ten unimolecular and bimolecular rate constants, we may try to choose the six constants $c_{1\mu}$ and $c_{2\mu}$ so that all δk 's except δk_{10} and δk_{20} vanish. If such c 's can be found, then they will determine associated shifts in fluxes δk_{10} and δk_{20} and concentrations δx_1 and δx_2 . Under these circumstances, the transformed kinetic equations will read

$$\begin{aligned} \dot{\bar{x}}_1 &= \bar{k}_{10} + k_{11} \bar{x}_1 + k_{12} \bar{x}_2 + k_{111} \bar{x}_1 \bar{x}_1 + k_{112} \bar{x}_1 \bar{x}_2 + k_{122} \bar{x}_2 \bar{x}_2 \\ \dot{\bar{x}}_2 &= \bar{k}_{20} + k_{21} \bar{x}_1 + k_{22} \bar{x}_2 + k_{211} \bar{x}_1 \bar{x}_1 + k_{212} \bar{x}_1 \bar{x}_2 + k_{222} \bar{x}_2 \bar{x}_2. \end{aligned} \quad (7.10)$$

Here, the k 's without overbars have the value taken on at the original ambient temperature, the change in the actual temperature-dependent k 's having been absorbed in the indicated changes in x_1, x_2, k_{10}, k_{20} indicated by overbars. When the \bar{x}_i are expressed in terms of the untransformed variables, the \bar{x}_i are seen to be lumped concentration variables if c_{12}, c_{21} , respectively, are nonzero. Otherwise, \bar{x}_1, \bar{x}_2 are simply altered values of x_1, x_2 .

Clearly, all this will only be possible in special cases — cases which may be determined using this linear analysis. When the linear analysis using infinitesimal

transformations establishes that compensation is possible, the corresponding finite transformations may be used to determine the shifts in fluxes and concentrations required to compensate for finite temperature-driven shifts in rate constants.

When this is possible, eqs. (7.10) state that the reaction with altered fluxes and concentration variables will proceed with the same unimolecular and bimolecular rate constants as did the original reaction at ambient temperature. If c_{12} and c_{21} are zero, one will have been able to accomplish this simply by changing fluxes and real world concentrations.

We also call attention to the fact that in the general case the determination of lumpings that will eliminate intermediates from consideration also begins with the determination of an appropriate infinitesimal transformation by specifying an appropriate linear combination of base generators. Once this has been determined — by solving a set of linear equations — one can determine the corresponding finite transformations. In proceeding from the infinitesimal to the finite transformations in these lumping analyses that fix a generator U , one may directly use the operator $\exp(aU)$ or a succession of different T 's, each involving one of the base generators in U and a particular choice of parameter that may be determined with the aid of table 2.2 of I or an extension of it that deals with a larger number of variables x_i and k_μ .

8. Invariant functions of kinetic coefficients

As the parameter a varies, the operators $\exp(aU)$ change the values of the kinetic coefficients k and the representative points in k space move along a definite path, as indicated in fig. 8.1 for a three-dimensional k space. The functional form of these paths is most usefully characterized by stating the functions $F(k)$ that are left invariant as the point moves along the path. Setting each $F(k)$ equal to a constant defines a surface in the space of kinetic coefficients, and the intersection of all these surfaces defines a line in this space — a path specified by the transformation. The constant value to be assigned to each $F(k)$ is determined by the initial values of the k 's. In the figure, it is supposed that both curves are determined by the same two generators U so that only the differing values of the constants C distinguishes them.

We now turn to the problem of determining the functions F . Let $F(k)$ be a function left invariant by the transformations $\exp(aU)$. Then, expanding the exponential, one has

$$\{1 + aU + (aU)^2/2 + \dots\} F = F. \quad (8.1)$$

The necessary and sufficient condition that this holds for all values of a is

$$UF = 0. \quad (8.2)$$

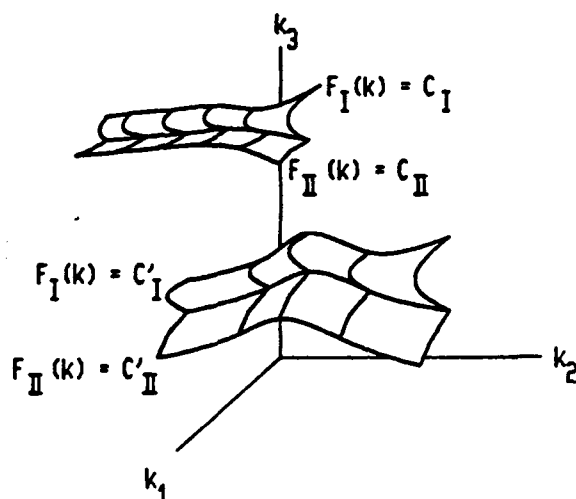


Fig. 8.1. Invariant surfaces and curves defined by invariant functions of rate constants. The functions F_I and F_{II} , when set equal to the constants C , here define two-dimensional surfaces in a three-dimensional space of rate coefficients. These surfaces intersect in a line. Changing the values of the constants C changes the surfaces and their intersection.

For a given U , this is a first-order partial differential equation for F . By the usual theory of such equations, it is equivalent to a set of first-order ordinary differential equations [9]

$$\frac{\delta k_{10}}{g_{10}} = \frac{\delta k_{20}}{g_{20}} \dots = \frac{\delta k_{222}}{g_{222}} \quad (8.3)$$

Consider, for example, the case of the transformation with generator

$$U_{11} = x_1 \partial/\partial x_1 + k_{10} \partial/\partial k_{10} + k_{12} \partial/\partial k_{12} - k_{111} \partial/\partial k_{111} \\ - k_{122} \partial/\partial k_{122} - k_{21} \partial/\partial k_{21} - 2k_{211} \partial/\partial k_{211} - k_{212} \partial/\partial k_{212} \quad (8.4)$$

Here, the equations (8.3) have as solutions a basic set of invariant functions

$$k_{10}/k_{12}, \quad k_{11}, \quad k_{111} \cdot k_{12}, \quad k_{112}, \quad k_{122}/k_{12} \\ k_{20}, \quad k_{21} \cdot k_{12}, \quad k_{22}, \quad k_{211} \cdot k_{12}^2, \quad k_{212} \cdot k_{12}, \quad k_{222} \quad (8.5)$$

Any function of these base functions is, of course, also an invariant function.

Table 8.1(a)
Invariants of the transformations

T_{10}	k_{111}	k_{112}	k_{122}	k_{211}	k_{212}	k_{222}
T_{11}	k_{11}	k_{20}	k_{22}	k_{112}	k_{222}	k_{10}/k_{12}
T_{12}	k_{20}	k_{21}	k_{211}	$k_{11} + k_{22}$	$k_{10}k_{21} + k_{20}k_{221}$	k_{10}/k_{12}
T_{111}	k_{10}	k_{12}	k_{20}	k_{21}	k_{22}	$2k_{111} + k_{212}$
T_{112}	k_{10}	k_{20}	k_{21}	k_{22}	k_{211}	k_{122}
T_{122}	k_{10}	k_{11}	k_{20}	k_{21}	k_{22}	k_{222}
					k_{22}	k_{111}

Table 8.1(b)
Invariants of the transformations

T_{10}	$(k_{11})^2 - 4k_{10}k_{111}$	$(k_{21})^2 - 4k_{20}k_{211}$	$2k_{12}k_{111} - k_{11}k_{112}$	$2k_{211}k_{22} - k_{21}k_{212}$	$k_{212}k_{12} - k_{112}k_{22}$
T_{11}	$k_{111}k_{12}$	k_{122}/k_{12}	$k_{21}k_{12}$	$k_{211}/(k_{12})^2$	k_{212}/k_{12}
T_{12}	$(k_{21})^2 - k_{211}k_{222}$	$k_{11}k_{22} - k_{12}k_{21}$	$k_{11}k_{211} - k_{21}k_{111}$	$k_{111}k_{212} - k_{112}k_{211}$	
T_{111}	k_{212}	k_{222}	$k_{10}k_{112} - k_{11}k_{12}$	$k_{11}k_{21} - 2k_{10}k_{211}$	$(k_{11})^2 - 4k_{10}k_{111}$
T_{112}	$2k_{10}k_{122} - (k_{12})^2$	$k_{111} + k_{212}$	$k_{20}k_{112} - k_{11}k_{22}$	$k_{11}k_{21} - k_{20}k_{111}$	$k_{10}k_{12} - k_{20}k_{12}$
T_{122}	k_{211}	k_{212}	$k_{12}k_{21} - k_{20}k_{112}$	$k_{112} + k_{222}$	$k_{11}k_{12} - 2k_{10}k_{122}$

Table 8.1(c)
Invariants of the transformations

T_{20}	k_{111}	k_{112}	k_{122}	k_{211}	k_{212}	k_{222}
T_{21}	k_{10}	k_{12}	k_{122}	$k_{11} + k_{22}$	$(k_{20}k_{12} + k_{10}k_{112})$	$2k_{222} + k_{112}$
T_{22}	k_{11}	k_{10}	k_{22}	k_{111}	k_{212}	k_{20}/k_{21}
T_{211}	k_{20}	k_{22}	k_{10}	k_{12}	k_{11}	k_{222}
T_{212}	k_{10}	k_{20}	k_{12}	k_{11}	k_{122}	k_{111}
T_{222}	k_{20}	k_{21}	k_{10}	k_{12}	k_{11}	k_{211}

Table 8.1(d)
Invariants of the transformations

T_{20}	$(k_{22})^2 - 4k_{20}k_{222}$	$(k_{12})^2 - 4k_{10}k_{122}$	$2k_{21}k_{222} - k_{22}k_{221}$	$2k_{122} - k_{12}k_{121}$	$k_{121}k_{21} - k_{212}k_{11}$
T_{21}	$(k_{112})^2 - k_{1122}k_{111}$	$k_{11}k_{22} - k_{12}k_{21}$	$k_{22}k_{122} - k_{12}k_{222}$	$k_{222}k_{111} - k_{212}k_{122}$	k_{112}/k_{21}
T_{22}	$k_{222}k_{21}$	k_{211}/k_{21}	$k_{12}k_{21}$	$k_{122}/(k_{21})^2$	k_{112}/k_{21}
T_{211}	k_{122}	k_{112}	$k_{12}k_{21} - k_{10}k_{212}$	$k_{212} + k_{111}$	$k_{22}k_{21} - 2k_{21}k_{11} + 2k_{10}k_{211}$
T_{212}	$2k_{20}k_{222} - (k_{21})^2$	$k_{222} + k_{112}$	$k_{10}k_{212} - k_{22}k_{21}$	$k_{21}k_{12} - k_{10}k_{222}$	$k_{20}k_{22} - k_{10}k_{21}$
T_{222}	k_{112}	k_{111}	$k_{20}k_{212} - k_{22}k_{21}$	$k_{22}k_{12} - 2k_{20}k_{122}$	$(k_{22})^2 - 4k_{20}k_{222}$

The reader will note on inspecting table 2.2 of I that the invariant functions (8.5) can also be constructed by eliminating the group parameter a from the finite transformations. If it should happen that k_{12} were zero, one would avoid introducing k_{12} by combining the transformed k 's in a different manner than indicated in (8.5).

In table 8.1, we list a basis of independent functions $F(k)$ left invariant by each of the generators in table 2.1 of I. Any two sets of values of the kinetic coefficients that give the same values for one or more of these sets of functions will yield reaction systems whose global behaviour is qualitatively the same in the sense defined in section 7. A set of eleven such basis functions $F(k)$ may be similarly determined for any linear combination of generators one chooses.

As an example of the utilization of these functions, we consider the functions determined by the translation operator $T_{10}(-a)T_{20}(-b) = T(-a, -b)$. This operator acts on (x_1, x_2) to give $(\bar{x}_1, \bar{x}_2) = (x_1 - a, x_2 - b)$. At the same time, it shifts a number of rate constants k_{μ} to \bar{k}_{μ} . $T(-a, -b)$ thereby determines homeomorphisms of x, k space that convert a given set of initial concentration values (x_1^0, x_2^0) (and running values (x_1, x_2)), and a given rate equation $\dot{x} = r(x, k)$ into a new set of concentrations obeying a new set of rate equations. For each value of a, b , the new initial concentrations $(\bar{x}_1^0, \bar{x}_2^0)$ evolve along a phase trajectory $(\bar{x}_1(t), \bar{x}_2(t))$ topologically equivalent to that of the initial phase trajectory $(x_1(t), x_2(t))$. Thus, by acting on a system with initial concentrations evolving along a phase trajectory of given topology, the transformation converts it into a two-parameter family of initial concentrations and phase trajectories of identical topology but belonging to different rate equations. (Any of the values (x_1, x_2) on the initial trajectory can of course be considered initial concentrations.) Inserting the initial values of the k_{μ} into the functions of table 8.1, one obtains initial values of the invariant functions. Setting the corresponding functions of the \bar{k}_{μ} equal to these initial values, one obtains the equations that determine the relations among the \bar{k}_{μ} that must subsist to ensure that the altered kinetic equations should have topologically identical trajectories originating from the transformed concentrations.

9. Group properties

So far, we have not dealt with important questions concerning the totality of transformations in table 2.2 of I. For example, are the different one-parameter groups of transformations in the table all subgroups of a single many-parameter group? Are there other time-independent transformations with generators quadratic in x , which will also leave the kinetic equations (2.1) invariant?

The first of these questions is also the logically prior one, because if the transformations do not together comprise a group, it can be shown that they give rise to further transformations which leave eqs. (1.1) invariant. Now, for the transformations to be those of a many-parameter group it is necessary and sufficient that their generators close under commutation:

$$[U_i, U_j] = \sum c_{ij}^k U_k. \quad (9.1)$$

In the previous paper I, we established that the commutation relations of invariance generators which leave the k subspace invariant are the same as the commutation relations of the full generators which act in the space of k and x . (That is to say, the structure constants c_{ij}^k are the same in both instances.) Because we have chosen the functions $h_i(x)$ to be independent of the k 's, it is also true that the commutation relations of that portion of the generators which acts on the x 's — the $\hbar \cdot \nabla_x$ — are also the same as the commutation relations of the full generators. This enables us to use Lie's classification of all the transformation groups of the plane (here the plane of x_1, x_2) to determine all possible Lie groups obtainable from the generators in table 2.1 of I. These are set forth in table 9.1.

Table 9.1

U's that generate many-parameter Lie groups

I.	$U_{10}, U_{20}, U_{11}, U_{12}, U_{21}, U_{22}, U_{111} + U_{212}, U_{222} + U_{112}$ (projective group of the plane [20])
II.	(i) $U_{10}, U_{20}, U_{11}, U_{21}, U_{22}, U_{211}, U_{111} + U_{212}$ (ii) $U_{20}, U_{10}, U_{22}, U_{12}, U_{11}, U_{122}, U_{222} + U_{112}$
III.	(i) $U_{10}, U_{20}, U_{11} + U_{22}, U_{21}, U_{211}, U_{111} + 2U_{212}$ (ii) $U_{20}, U_{10}, U_{22} + U_{11}, U_{12}, U_{122}, U_{222} + 2U_{112}$
IV.	(i) $U_{10}, U_{11}, U_{12}, U_{112}, U_{20}, U_{22}$ (ii) $U_{20}, U_{22}, U_{21}, U_{212}, U_{10}, U_{11}$
V.	(i) $U_{10}, U_{12}, U_{122}, U_{20}, U_{22}, 2U_{11} + U_{122}$ (ii) $U_{20}, U_{21}, U_{211}, U_{10}, U_{11}, 2U_{22} + U_{211}$
VI.	(i) $U_{10}, U_{20}, U_{11}, U_{22}, U_{21}, U_{211}$ (ii) $U_{20}, U_{10}, U_{22}, U_{11}, U_{12}, U_{122}$
VII.	$U_{10}, U_{20}, U_{11}, U_{12}, U_{21}, U_{22}$ (general linear group of the plane [20])
VIII.	$U_{10}, U_{20}, U_{12}, U_{21}, U_{11} - U_{22}$ (special linear group of the plane [20])
IX.	(i) $U_{10}, U_{11}, U_{22}, U_{122} + U_{112}$ (ii) $U_{20}, U_{22}, U_{11}, U_{211} + U_{212}$
X.	(i) $U_{10}, 2U_{11} + U_{22}, U_{111} + U_{212}$ (ii) $U_{20}, 2U_{22} + U_{11}, U_{222} + U_{112}$
XI.	(i) $U_{10}, U_{11}, U_{12}, U_{122}$ (ii) $U_{20}, U_{22}, U_{21}, U_{211}$
XII.	(i) U_{10}, U_{12}, U_{122} (ii) U_{20}, U_{21}, U_{211}
XIII.	(i) U_{10}, U_{11}, U_{111} (ii) U_{20}, U_{22}, U_{222} (group of the line [20])

Note: Many of the groups whose generators are listed above contain subgroups not listed, e.g. in XIII (i), U_{10} and U_{11} generate a two-parameter group.

It will be noted that no one of the many-parameter groups in this table contains all the generators in table 2.1 of I. The largest group is the first listed, a ten-parameter group that is a form of the projective group of the plane. If one takes the commutators of the generators in this group with the remaining linearly independent generators available from table 2.1 of I, then one obtains new generators not in table 2.1. However, in the generators the h_i are of third degree in x . No further linearly independent generators exist in which the h are of less than third degree and g is nonzero.

10. Errors in finite transformations resulting from use of approximate generators

The generators used in section 5 to approximately linearize the Lotka–Volterra equations are typical generators in the sense that they are generators of transformations that only approximately leave invariant a set of kinetic equations. Expanding the finite transformation operator $\exp(aU)$ in powers of the group parameter a , one sees that as a consequence one would have to expect that the effect of $\exp(aU)$ on the differential equation, its solutions, and functions of its solutions, would only be accurate through $O(ax^2)$. In particular, eq. (5.2) is linearized only through $O(y^2)$. However, one is interested in having the transformation $\exp(aU)$ act at every point on a given solution curve – not just near the origin.

In sections 3 and 5, we have used critical points in phase space as origins of coordinates. One can just as well choose a point on or near a trajectory as the origin and thereby ensure that in the region of such a point, the error in the coefficients $h_i(x)$ in U is minimal. This allows one to determine trajectories in the region of any point P that are accurate through second order in displacements from P . If, using P as origin, one proceeds as in sections 3, 5 and transforms the system of interest into a system with known analytic solutions, one can use the inverse transformation to obtain analytic approximations to trajectories in the region of P . From a more general standpoint, expansions about P will allow accurate investigations of solution behaviour near P when one varies k 's.

To illustrate the method, we use it to improve the approximate Lotka–Volterra trajectory obtained in section 5. There, the analytic reference solution was obtained by transforming away the quadratic terms in the rate equations using an operator $\exp(aU)$. Since the group generators are accurate to $O(y^3)$, this gave a set of rate equations linear to $O(y^3)$, the origin being the singular point. The linear equations were solved, and their solution transformed into an approximate solution of the Lotka–Volterra equation (5.3) by action of $\exp(-aU)$.

To improve the solutions obtained in this way, one may proceed as follows:

- (i) Determine the general form of the generator of the transformation that linearizes the nonlinear equations in the region of a point P on the actual trajectory of interest – e.g. the point whose coordinates are initial values of the species concentrations.

- (ii) Determine the finite transformation that carries out the linearization.
- (iii) Obtain and solve the linearized equations.
- (iv) Transform the solution of the linearized equation into the required solution of the nonlinear equation.

Let the new center of expansion of (5.2) be at a point P with coordinates (α, β) and define

$$y_1^\alpha = y_1 - \alpha, \quad y_2^\beta = y_2 - \beta \quad (10.1a)$$

and

$$T_{10}(-\alpha, -\beta) = \exp(-\alpha U_{10} - \beta U_{20}), \quad y^{\alpha\beta} = T_{10}(-\alpha, -\beta)y = (y_1^\alpha, y_2^\beta). \quad (10.1b)$$

The action of $T_{10}(-\alpha, -\beta)$ on eqs. (5.2) gives

$$\begin{aligned} dy_1^\alpha/dt &= k_{10}^{\alpha\beta} + k_{11}^{\alpha\beta} y_1^\alpha + k_{12}^{\alpha\beta} y_2^\beta + k_{112}^{\alpha\beta} y_1^\alpha y_2^\beta \\ dy_2^\beta/dt &= k_{20}^{\alpha\beta} + k_{21}^{\alpha\beta} y_1^\alpha + k_{22}^{\alpha\beta} y_2^\beta + k_{212}^{\alpha\beta} y_1^\alpha y_2^\beta, \end{aligned} \quad (10.2a)$$

where

$$\begin{aligned} k_{10}^{\alpha\beta} &= \alpha\beta k_{112} + \beta k_{12}, & k_{11}^{\alpha\beta} &= \beta k_{112}, \\ k_{12}^{\alpha\beta} &= k_{12} + \alpha k_{112}, & k_{112}^{\alpha\beta} &= k_{112}, \\ k_{20}^{\alpha\beta} &= \alpha\beta k_{212} + \alpha k_{21}, & k_{21}^{\alpha\beta} &= k_{21} + \beta k_{212}, \\ k_{22}^{\alpha\beta} &= \alpha k_{21}, & k_{212}^{\alpha\beta} &= k_{212}. \end{aligned} \quad (10.2b)$$

We seek an invariance generator

$$U = h(y^{\alpha\beta}) \cdot \nabla_{y^{\alpha\beta}} + g \cdot \nabla_{k^{\alpha\beta}} \quad (10.3)$$

and a value of a such that $\exp(aU)$ acts on $y^{\alpha\beta}$ and $k^{\alpha\beta}$ to transform the $k_{112}^{\alpha\beta}$ and $k_{212}^{\alpha\beta}$ terms to zero, leaving only terms of $O((y^{\alpha\beta})^0)$, $O(y^{\alpha\beta})$, and $O((y^{\alpha\beta})^3)$ and higher. One may suppose that such a generator is of the form $\Sigma c_\mu U_\mu$. We first determine the c_μ that would be required if the nonlinearity were infinitesimal. To do this, we multiply $k_{112}^{\alpha\beta}$ and $k_{212}^{\alpha\beta}$ by an infinitesimal ϵ and determine the c_μ by requiring that $(1 + \delta aU)$ annihilate $\epsilon k_{112}^{\alpha\beta}$ and $\epsilon k_{212}^{\alpha\beta}$ while leaving $k_{111}^{\alpha\beta}$, $k_{122}^{\alpha\beta}$, $k_{211}^{\alpha\beta}$ and $k_{222}^{\alpha\beta}$ all zero.

Inspecting table 2.1 of I, one finds that in the sum one need only consider the six generators U_{ij} that generate nonlinear transformations of the concentrations. Considered as functions of y_1^α, y_1^β , all these generators vanish at $y_1^\alpha = 0 = y_2^\beta$. It follows that $\bar{y}_1^\alpha, \bar{y}_2^\beta$ also vanish at the origin, which is thus an invariant point of the transformation. Table 2.2 of I shows the transformed rate constants \bar{k}_{i0} and \bar{k}_{ij} depend linearly on both group parameters and rate constants. Consequently, $\exp(U)$ has the same effect on the k_{0i} and k_{ij} as does $(1 + U)$, so that setting $\epsilon = \delta a$ allows one to use $(1 + U)$ to obtain the same linearized equation as would be obtained using $\exp(U)$. It cannot, however, be concluded that $(1 + U)$ generally acts on the concentration variables to give transformed variables that are good approximations to those obtained by the action of $\exp(U)$.

For $(1 + \delta a U)$ to kill $\epsilon k_{ij}^{\alpha\beta}$, the c_μ must satisfy the following set of linear equations:

$$\begin{aligned}
 0 &= \delta a (c_{111} k_{11}^{\alpha\beta} + c_{112} k_{21}^{\alpha\beta} - c_{211} k_{12}^{\alpha\beta}) \\
 -\epsilon k_{112}^{\alpha\beta} &= \delta a (c_{111} 2k_{12}^{\alpha\beta} + c_{112} k_{22}^{\alpha\beta} + c_{122} 2k_{21}^{\alpha\beta} - c_{212} k_{12}^{\alpha\beta}) \\
 0 &= \delta a (c_{112} k_{12}^{\alpha\beta} + c_{122} (2k_{22}^{\alpha\beta} - c_{222} k_{12}^{\alpha\beta})) \\
 0 &= \delta a (-c_{111} k_{21}^{\alpha\beta} + c_{211} (2k_{11}^{\alpha\beta} - k_{22}^{\alpha\beta}) + c_{212} k_{21}^{\alpha\beta}) \\
 -\epsilon k_{212}^{\alpha\beta} &= \delta a (-c_{112} k_{21}^{\alpha\beta} + c_{211} 2k_{12}^{\alpha\beta} + c_{212} k_{11}^{\alpha\beta} + c_{222} 2k_{21}^{\alpha\beta}) \\
 0 &= \delta a (-c_{122} k_{21}^{\alpha\beta} + c_{212} k_{12}^{\alpha\beta} + c_{222} k_{22}^{\alpha\beta})
 \end{aligned} \tag{10.4}$$

To further particularize the discussion, we approximate a trajectory of the Lotka-Volterra equations (5.1) through the point (0.922, -0.491). Translating the origin to this point, the Lotka-Volterra equations become

$$\begin{aligned}
 \dot{y}_1^\alpha &= 0.9437 + 0.491 y_1^\alpha - 1.922 y_2^\beta - y_1^\alpha y_2^\beta \\
 \dot{y}_2^\beta &= 0.4693 + 0.509 y_1^\alpha + 0.922 y_2^\beta + y_1^\alpha y_2^\beta.
 \end{aligned} \tag{10.5}$$

To linearize these, we first use (10.4) to determine the parameters a_{ijk} in the linearizing operator $1 + \sum a_{ijk} U_{ijk}$, and find them to be

$$\begin{aligned}
 a_{111} &= -0.3289, \quad a_{112} = -0.1067, \quad a_{122} = 0.4829 \\
 a_{211} &= 0.1123, \quad a_{212} = -0.3422, \quad a_{222} = -0.4467.
 \end{aligned} \tag{10.6}$$

The approximate y linearized equations, obtained using $1 + \sum a_{ijk} U_{ijk}$, are

$$\begin{aligned}\dot{\bar{y}}_1^\alpha &= 0.9437 + 0.1170 \bar{y}_1^\alpha - 1.0511 \bar{y}_2^\beta + O(y^3) \\ \dot{\bar{y}}_2^\beta &= 0.4693 + 0.7161 \bar{y}_1^\alpha + 0.1615 \bar{y}_2^\beta + O(y^3).\end{aligned}\tag{10.7}$$

If one writes the finite transformation T in the form

$$T = T_{222} [T_{212} [T_{211} [T_{122} [T_{112} [T_{111}]]]]], \tag{10.8}$$

one finds that T linearizes (10.5), yielding (10.7), when the group parameters are

$$\begin{aligned}a_{111} &= -0.0123, \quad a_{112} = -0.7473, \quad a_{122} = 1.6794 \\ a_{211} &= 0.2792, \quad a_{212} = -0.6817, \quad a_{222} = -0.1248.\end{aligned}\tag{10.9}$$

There are several ways to obtain these values. We calculated them by taking advantage of the fact that when the k_{i0} vanish, T acts linearly on the k_{ijj} , and so began with initial approximations to the a 's which we obtained by solving (10.4). We then simultaneously increased k_{10} and k_{20} in five stages. At each stage, the a 's that zeroed the k_{ijj} to 1 part in 10^4 were determined by Newton's method. This required two steps at each stage, and yielded final values of the a 's that zero the k_{ijj} to within 1 part in 10^5 .

The solution of (10.7) passing through $\bar{y}^\alpha = 0 = \bar{y}^\beta$ at $t = 0$ obtained on neglecting terms $O(y^3)$ is

$$\begin{aligned}\bar{y}_1^\alpha &= -0.8349 + 0.8349 \cos(0.8673 t) \\ &\quad + 0.9549 \sin(0.8673 t) \exp(0.1384 t) \\ \bar{y}_2^\beta &= 0.8049 + (-0.8049 \cos(0.8673 t) \\ &\quad + 0.6695 \sin(0.8673 t) \exp(0.1384 t)).\end{aligned}\tag{10.10}$$

Acting on $(\bar{y}_1^\alpha, \bar{y}_2^\beta)$, the inverse transformation T^{-1} gives (y_1^α, y_2^β) . In fig. 10.1, the resulting phase trajectory is compared with the exact trajectory and with the trajectory generated by dropping the quadratic terms in (10.5), and then solving the resulting linear equation. The errors in the trajectory obtained by transformation arise via third-order errors in the linearized equations. The errors in the other trajectory arise from second-order errors in the linearized equations.

It should be noted that the phase trajectory of (5.2) passing through the point P with coordinates $(y_1, y_2) = (0.922, -0.491)$ is a closed curve. However, when the translated equation (10.5) is linearized by dropping its bimolecular terms,

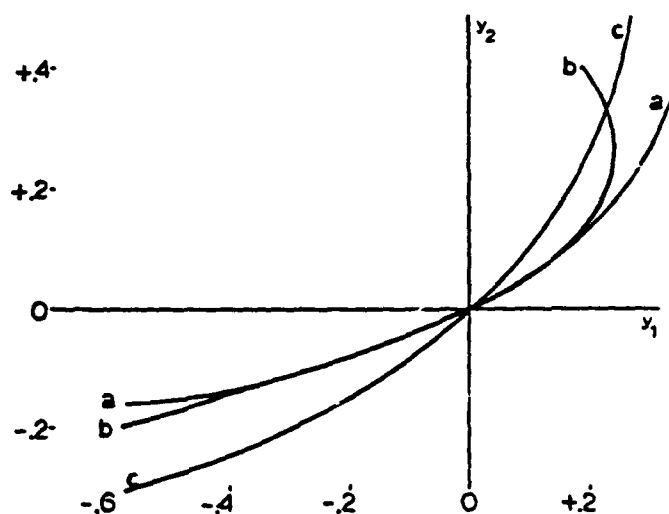


Fig. 10.1. Regional approximation to a phase trajectory of the Lotka – Volterra equation. Curve *a* is an exact trajectory of (10.5). Curve *b* is its regional approximation defined by (10.8, 10.9, 10.10). Curve *c* is the approximation to curve *a* determined by the usual linearization of (10.5).

all its phase curves are open ones. The linearization is not an invariant one in our generalized sense (cf. I), and has as a consequence not left the topology of its phase curves invariant. The same is true of the regional linearization method: (10.7) has only open curves for phase trajectories because our generators are insufficiently accurate to ensure that the approximate linearization carried out by T is a sufficiently good approximation to an invariance transformation. The open phase trajectories of (10.7) are then of course mapped into open phase trajectories by the transformation inverse to (10.8) because the transformation is a diffeomorphism. These topological errors could of course have been avoided had we linearized equations (5.2) in the way we did in section 5, and then translated the resulting equations to the new origin. This, however, makes it more difficult to obtain a close approximation to the phase curves at points far from the singular point at the origin. The method illustrated here is designed for that purpose.

11. Higher approximations to generators

All our considerations so far have involved generators obtained by quadratic approximation. In this section, we will determine higher approximations to the generators and investigate the ways in which their use modifies results obtained from the quadratic approximation. It will be remembered that the quadratic approxi-

mation to the U was obtained by solving eqs. (2.8a), (2.8b) together with the approximation to (2.8c) obtained by setting $U^{(2)}$ to zero. We begin this section by relaxing the approximation that $U^{(2)} = 0$ in (2.8c), and thereby solve the full set of equations implied by (2.8a, b, c). Inspecting (2.8), one sees that this completely determines the k terms in the U . Thus, the approximation we are about to discuss fixes the g 's and therefore for each U completely determines the transformation of the kinetic coefficients carried out by $\exp(aU)$.

We start with an example and determine the modifications to the U_{122} of table 2.1 of I that one obtains by removing the approximation $U^{(2)} = 0$ when solving (2.8a, b, c) of I. Equations (2.8a, b) are not altered and one obtains from (2.8c) the six determining equations

$$\begin{aligned}
 (g_{111}) - 3k_{10}h_{1111} - k_{20}h_{1112} &= 0 \\
 (g_{112} - 2k_{21}) - 2k_{10}h_{1112} - 2k_{20}h_{1122} &= 0 \\
 (g_{122} + k_{11} - 2k_{22}) - k_{10}h_{1122} - 3k_{20}h_{1222} &= 0 \\
 (g_{211}) - 3k_{10}h_{2111} - k_{20}h_{2112} &= 0 \\
 (g_{212}) - 2k_{10}h_{2112} - 2k_{20}h_{2122} &= 0 \\
 (g_{222} + k_{21}) - k_{10}h_{2122} - 3k_{20}h_{2222} &= 0.
 \end{aligned} \tag{11.1}$$

On setting $U^{(2)} = 0$, the terms in parentheses remain and are the terms used previously to determine the $U^{(-1)} + U^{(0)} + U^{(1)}$ approximation to U . To obtain corrections to the resulting U_{122} , one transfers these terms to the right-hand side of the equations and solves the resulting inhomogenous equations for the h_{ijkl} . The three equations for the h_{1jkl} and the three for the h_{2jkl} are independent and each set is of rank 3 if neither k_{10} nor k_{20} vanish. Consider this case first. Solving the equations, one finds that they yield the following U :

$$\begin{aligned}
 U &= U_{122} + (-K^3 x_1^3 + 3K^2 x_1^2 x_2 - 3K x_1 x_2^2 + x_2^3) \\
 &\quad \times (e_1 \partial/\partial x_1 + e_2 \partial/\partial x_2),
 \end{aligned} \tag{11.2}$$

where $K = k_{20}/k_{10}$. Here, e_1 and e_2 are arbitrary parameters. One may in fact re-express (11.2) in the form

$$U = U_{122} + e_1 U_{e_1} + e_2 U_{e_2} \tag{11.3}$$

As U_{122} is reclaimed on setting e_1, e_2 to zero, U_{122} is itself a solution of the full set of equations $W_i = 0$. Thus, U_{122} is one degree more accurate than might have been expected. It will also be noted that the operators U_{e_1} and U_{e_2} act only on x_1, x_2 and not upon the rate coefficients k . They are consequently of no interest in the context of this paper.

Next, consider the case $k_{10} = k_{20} = 0$. It is evident that each of the h_{ijk} may then be chosen arbitrarily, so that one obtains an eight-parameter family of generators:

$$U = U_{122} + \sum h_{ijk} x_j x_k x_i \partial / \partial x_i. \quad (11.4)$$

As in the previous case, the additional generators have no effect upon the rate coefficients.

Next, consider the situation where k_{20} vanishes, while k_{10} does not. Then one finds

$$U = U_{122} + h_{1222} x_2^3 \partial / \partial x_1 + h_{2222} x_2^3 \partial / \partial x_2. \quad (11.5)$$

When k_{10} vanishes and k_{20} does not, one finds

$$U = U_{122} + h_{1111} x_1^3 \partial / \partial x_1 + h_{2111} x_1^3 \partial / \partial x_2. \quad (11.6)$$

In both cases, the h 's are arbitrary and are coefficients of new generators that have no effect on the rate constants. In short, in order to obtain corrections to U_{122} it is necessary to move on to eq. (2.8d) of I.

This discussion of "corrections" to U_{122} applies to the other U_{ijk} in a parallel manner. The terms in (11.1) not contained in parentheses are the same in each case. The terms contained in parentheses are different in each case, but vanish in the original approximation. Thus, the generators listed in table 2.1 of I and the finite transformations in table 2.2 of I are all unchanged when eqs. (2.8a, b, c) of I are solved in toto.

We next investigate the modifications of the $U^{(2)}$ that are required in order to satisfy (2.8d) of I. Equation (2.8d) may be written in matrix form as

$$0 = G^{(2)} H^{(2)} + G^{(1)} H^{(3)} + G^{(0)} H^{(4)} = (GH)^{(4)}. \quad (11.7)$$

Here, $G^{(n)}$ is a matrix whose entries contain $g^{(n)}$ coefficients and $H^{(n)}$ is a vector of $h^{(n)}$ coefficients. The product $(GH)^{(4)}$ is of the form

$$(GH)^{(4)} = \begin{bmatrix} [g^{(0)}] & [0] & [0] \\ [0] & [g^{(1)}] & [0] \\ [0] & [0] & [g^{(2)}] \end{bmatrix} \begin{bmatrix} [h^{(4)}] \\ [h^{(3)}] \\ [h^{(2)}] \end{bmatrix} \quad (11.8)$$

From this, it is evident that on insertion into (11.8) of the $h^{(2)}$ and $h^{(3)}$ calculated by setting to zero the lower order w , one obtains a set of equations which determine the $h^{(4)}$ without modifying the lower order $h^{(n)}$. It follows that the functions $g(k)$ in the generators obtained by solving (2.8a, b, c) of I are exact. Thus, the invariant functions listed in table 8.1 are exact.

If one wishes to use transformations whose generators are linear combinations of those listed in table 2.1 of I, it becomes necessary to integrate eqs. (8.3) to determine the corresponding invariant functions of the rate constants. These also will remain unaltered by all further improvements in the generators obtained by solving eqs. (2.8) of I in higher orders of approximation.

An interesting property of the higher order approximations to the U 's is worth noting. Even when a set of U_r in table 2.1 of I close under commutation, it will not generally be true that the corresponding set of improved generators will close under commutation. The commutators will generally contain terms of higher degree in x than the original generators. However, one may write

$$U_n = {}^kU_r + {}^xU_r, \quad (11.9)$$

where kU_r acts only on the kinetic coefficients and xU acts only on the species concentrations. If the kU_r close under commutation, then the theorem of ref. [1] of I establishes that the U_r will obey the same commutation relations as the kU_r when they satisfy (2.8) of I exactly. Any failure of the approximate generators to obey these commutation relations is thus an artifact of approximation.

Finally, we consider the general problem of obtaining arbitrarily high-order approximations to a generator U . Referring back to eqs. (2.8) of I, one sees that the contribution to U of order $p+1$ in x is obtained from the contributions of order p and $p-1$ by solving linear equations exactly analogous to those depicted in (11.8) above. As in the case of the example of eqs. (11.1), one obtains solutions corresponding to generators with g vanishing as well as the desired improvement $U^{(p+1)}$ to the U of interest. This $U^{(p+1)}$ can then be used together with $U^{(p)}$ to obtain $U^{(p+2)}$ in an analogous fashion.

12. Conclusions

This paper has utilized basic methods of the theory of Lie groups admitted by ordinary differential equations to determine large-scale global mappings connecting systems with differing rate constants.

As we have illustrated, a key consequence of such large changes is their effect upon the topology of the phase trajectories of a system. As we knew that time-independent transformations of species concentrations and rate constants could preserve the topology of phase portraits if the transformations were sufficiently

restricted, in this paper we investigated time-independent transformations whose generators are analytic in the rate constants and approximated as analytic in the concentrations. This is more than sufficient to force the transformations to be local diffeomorphisms of the entire system space — the space of all real values of the concentrations and rate constants. By also restricting the range of the group parameter where necessary, we have ensured that all finite transformations are diffeomorphisms of the space of real x, k . In addition, because the generators are so chosen that the space of rate constants is an invariant subspace, the topology of trajectories in concentration space is preserved by the transformations. This has allowed us to determine the one-parameter groups of changes in rate constants for which the phase trajectories are qualitatively insensitive in a well defined topological sense. As we have been able to exactly determine the changes in rate constants that preserve the topology of these phase portraits, it is possible to give a quantitative treatment of these changes in rate constants without further elaboration.

Because the determining equations for the group generators could be solved algorithmically, we have been able to systematically determine all one-parameter transformation groups satisfying the imposed conditions.

We are not the first to realize the importance of topological considerations in chemical kinetics: we particularly call attention to the work of Bruce Clark and his coworkers [10], and to the work of Martin Feinberg [11].

Our work differs from that of these and other investigators because we have taken advantage of the fact that the process of determining the Lie generators of an invariance transformation can be made algorithmic. This now makes it possible to develop a systematic and general treatment of the consequences of large changes in rate constants upon the behaviour of kinetic systems.

We have not attempted to exactly determine the phase portraits themselves. There is a fundamental reason for this. Autonomous ordinary differential equations whose right-hand sides are analytic functions can have "chaotic" solutions. This has the consequence that the coefficients $h(x)$ in the generators U of this paper need not be analytic functions; they may, for example, be only infinitely differentiable functions. In practice, one may approximate infinitely differentiable functions by a series of analytic functions, but it would be a mistake to suppose that this approximation was of the same value in all regions of the phase space. Experience suggests that this, and related, mathematical complexity seldom expresses itself in the chaotic evolution of the reacting systems of common occurrence in the chemical laboratory and chemical industry. It may be of more common occurrence in biochemical systems. Whenever the evolution of a kinetic system is nonchaotic, the transformations introduced in this paper allow one to both qualitatively and quantitatively investigate the sensitivity of phase trajectories to gross changes in rate constants, and to determine those changes in rate constants which leave some quantitative property unchanged [12]. If the evolution is chaotic, further investigations are necessary.

In the interest of simplicity, we have also side-stepped three problems mathematically much less troublesome than that of chaotic evolution. We have not required that the group parameters a be so restricted so as to ensure that no "real world" concentration becomes negative. We have also not required that mass conservation be preserved when $T(a)$ acts on a kinetic system. There are no fundamental problems involved here; it is not difficult to impose the requirements in any particular case — the difficulty is simply that the variety of cases is immense and diverse. Finally, we have not dealt with problems that arise when many-parameter Lie groups, whose parameters are only restricted in range by the structural properties of the group, have further restrictions imposed by the requirement that the group action on a space of real variables yields only real variables. In our case, the difficulty appears when abstractly allowed parameter values carry points with finite coordinates to coordinates whose value is $\pm\infty$. A considerable simplification occurs if one proceeds as is done in the theory of projective transformations; this, however, changes the topology of the space of x, k and introduces conceptual elaborations that we consider to be inappropriate in an introductory work such as this.

A variety of applications can be envisioned for the time-independent transformations of this paper. Because so much of the analysis involves only linear algebra, the methods are applicable to systems involving many chemical species. Further applications to the linearization of kinetics and to lumping and control problems appear to hold particular promise. The methods we have introduced for determining the subspace of x, k containing phase space trajectories of a fixed topology are methods that are systematic and apply directly to systems involving an arbitrary number of reactants: they may be used to obtain a great deal of qualitative information about these systems. The use of the methods to obtain regional analytic approximations to solutions of nonlinear kinetic equations also appear promising.

We are currently extending Lie methods to reactions involving diffusion [13]. It is known that reaction-diffusion equations are invariant under a much larger class of transformations than those considered herein and in I; in the general case, it will be necessary to allow transformations that depend upon partial derivatives of arbitrary order [14].

Acknowledgements

The authors wish to thank Guang-Hui Xu and Gordon Ballentine for assistance with the computations and figures. We also wish to acknowledge the support of this research by the Air Force Office of Scientific Research.

References

- [1] C.E. Wulfman and H. Rabitz, *J. Math. Chem.* 3(1989)
- [2] F.C. Frank, *Biochim. Biophys. Acta* 11(1953)459.
- [3] A.R. Hochstim, *Origins of Life* 6(1975)317.
- [4] A.J. Lotka, *Elements of Physical Biology* (Williams and Wilkins, 1925).
- [5] V. Volterra, *Mem. Acad. Lincei* 2(1926)31; cf. also:
V. Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie* (Paris, 1931).
- [6] Cf., for example, H.T. Davis, *Introduction to Nonlinear Differential and Integral Equations* (U.S. Atomic Energy Commission, Washington, D.C., 1960) p. 102.
- [7] V.I. Arnold, *Ordinary Differential Equations*, trans. by R.A. Silverman (MIT Press, Cambridge, MA, 1973).
- [8] W.E. Boyce and R.C. DiPrima, *Elementary Differential Equations and Boundary Value Problems* (Wiley, New York, 1977) p. 406.
- [9] S. Lie, *Vorlesungen über Continuerliche Gruppen*, Abteilung III (Chelsea, New York, 1971).
- [10] (a) B.L. Clarke, *Advances in Chemical Physics*, ed. I. Prigogine and S.A. Rice, Vol. 43, (Wiley, New York, 1980) pp. 1–215;
(b) B.L. Clarke, *J. Chem. Phys.* 75(1981)4970.
- [11] M. Feinberg, *Dynamics and Modelling of Reactive Systems*, ed. W. Stewart, W.H. Ray and C. Conley (Academic Press, New York, 1980) pp. 59–129.
- [12] Cf. C. Wulfman and H. Rabitz, *J. Phys. Chem.* 90(1986) for a discussion of the determination of group generators that leave kinetic equations and additional functions or functionals invariant.
- [13] H. Rabitz and C. Wulfman, to be published.
- [14] C. Wulfman and Tai-ichi Shibuya, *Rev. Mex. de Física* 22(1973)171.
- [15] Cf., for example, J.E. Campbell, *Introductory Treatise on Lie's Theory of Finite Continuous Transformation Groups* (Chelsea, New York, 1966), reprint of 1903 edition.